Behavioral and Brain Sciences

cambridge.org/bbs

How linguistics learned to stop worrying and love the language models

Richard Futrell[a] and Kyle Mahowald[b]

[a]University of California Irvine, Irvine, CA, USA, and [b]The University of Texas at Austin, Austin, TX, USA
rfutrell@uci.edu
kyle@utexas.edu

Abstract

Language models (LMs) can produce fluent, grammatical text. Nonetheless, some maintain that language models don't really learn language and also, even if they did, that would not be informative for the study of human learning and processing. On the other side, there have been claims that the success of LMs obviates the need for studying linguistic theory and structure. We argue that both extremes are wrong. LMs can contribute to fundamental questions about linguistic structure, language processing, and learning. They force us to rethink arguments and ways of thinking that have been foundational in linguistics. While they do not replace linguistic structure and theory, they serve as model systems and working proofs of concept for gradient, usage-based approaches to language. We offer an optimistic take on the relationship between language models and linguistics.

Target Article

Cite this article: Futrell R and Mahowald K. (2026) How linguistics learned to stop worrying and love the language models. *Behavioral and Brain Sciences* **49**, e198: 1–80. doi:10.1017/S0140525X2510112X

Target Article Accepted: 25 November 2024 UTC
Target Article Manuscript Online: 24 July 2025 UTC
Commentaries Accepted: 28 May 2025 UTC

Keywords:
language models; linguistic theory; language learning; information theory; functional linguistics; statistical learning; neural networks

What is Open Peer Commentary? What follows on these pages is known as a Treatment, in which a significant and controversial Target Article is published along with Commentaries (p. 27) and an Authors' Response (p. 69). See bbsonline.org for more information.

Corresponding authors: Richard Futrell; Email: rfutrell@uci.edu; Kyle Mahowald; Email: kyle@utexas.edu

Authors contributed equally. Author order was determined alphabetically by last name.

© The Author(s), 2025. Published by Cambridge University Press

1. Introduction

It's 1968, and Norm and Claudette are having lunch. Norm is explaining his position that all human languages share deep underlying structure and has worked out careful theories showing how the surface forms of language can be derived from these underlying principles. Claudette, whose favorite movie is the recently released *2001: A Space Odyssey* and who particularly loves the HAL character, wants to make machines that could talk with us in any human language. Claudette asks Norm whether Norm thinks his theories could be useful for building such a system. Norm says he is interested in human language and the human mind, found HAL creepy, and isn't sure why Claudette is so interested in building chatbots or what good would come of that. Nonetheless, they both agree that it seems likely that, if Norm's theories are right (and he sure thinks they are!), they could be used to work out the fundamental rules and operations underlying human language in general – and that should, in principle, prove useful for building Claudette's linguistic machines. Claudette is very open to this possibility: all she wants is a machine that talks and understands. She doesn't really care how it happens. Norm and Claudette have very different goals, but they enjoy their conversations and are optimistic that they can both help each other.

Fast forward to 2025. Norm has worked for decades on a variety of diverse languages, developing sophisticated theories of linguistic structure. Claudette got more and more interested in engineering, amassing huge amounts of data, and training statistical models. Norm and Claudette slowly fell out of touch over the years. They are planning a reunion lunch: what should they talk about?

Just like Claudette always dreamed of, she now has machines that can talk with her in English, producing grammatical and seemingly sensible utterances. How relevant is it that the architecture of Claudette's machines seems to have nothing to do with the structure of language as identified by Norm and other linguists and cognitive scientists? Claudette might not care: she just wanted a system that worked, and this stuff works. On the other hand, what if Norm was right about the nature of language – does that mean the machines aren't actually as impressive as Claudette thinks, because they are relying on shallow pattern matching? Or are Claudette's machines evidence that Norm's theories were wrong? More generally, what insights do these machines have for the scientific study of language?

One approach Norm could take would be to dismiss Claudette's machines out of hand as irrelevant to human language, as they are not realistic models of the human brain, nor are they intended to be. Under this view, neural networks are no more relevant to linguistics than submarine engineering is to an ichthyologist – just because both submarines and fish can move underwater does not mean that you can learn much about one from studying the other.

This negative view takes several forms. One argument is that language models (yes, Claudette's machines are language models; henceforth LMs) are so fundamentally different in their architecture from people and have access to so much more data, that whatever they are doing is so different as to be totally irrelevant for humans (Chomsky et al., 2023; Fox & Katzir, 2024; Bolhuis et al., 2024). Some have denied that language models have actually learned, or ever

could learn, the putatively key properties of human language, and thus are either irrelevant to language science or do not challenge any existing ways of thinking (Lan et al., 2024; Fox & Katzir, 2024). A different and more subtle argument is that neural network sequence models could learn to approximate *anything*, so the fact that they seem to learn language (to the extent that they do) is uninformative (Rawski & Baumont, 2023; Moro et al., 2023; Chomsky, 2023; Chomsky et al., 2023; Collins, 2024; Bolhuis et al., 2024). Under this view, LM architectures are like epicycles, the computational technique used by Ptolemy to predict the motions of the planets and the sun in a model that placed the Earth at the center of the universe (de Santillana, 1955; Flynn, 2013). The problem with epicycles is that they can approximate any trajectory arbitrarily well (at the cost of great complexity), so the mere fact that they could be used to capture the motions of the planets with high accuracy tells us nothing about the real nature of that motion. Similarly, if neural networks are just general pattern approximators, then the fact that they can approximate language doesn't tell us anything about language.

We believe these negative views are short-sighted: language models *do* learn nontrivial aspects of linguistic structure, and they do give important insights that change how we should think about language. As language scientists, we ignore them at our peril.

An opposite approach is to dismiss traditional theories of linguistic structure, on the grounds that they turned out to be either useless or of negative value in developing the only known systems that can actually use language as humans do (Jelinek, 2004; Piantadosi, 2023). Under this view, decades of work on the structure of language turned out to be barking up the wrong tree, and the theory of language should be rebuilt on top of the only computational foundation that has been demonstrated to be able to produce and comprehend language to a human level: large neural networks. In our experience, this view is widespread in some engineering and application-focused communities.

This view is also short-sighted. It throws out hard-won analytical discoveries about the structure of language and leaves us adrift in terms of a scientific theory of language, without a clear way to approach the question of *why* human language is the way it is, or even a sense of what the interesting questions are. Moreover, language models are currently most successful in English and other languages with internet-scale data (Blasi et al., 2022). A more complete approach to the science of language will draw on the expertise of documentary linguists, sociolinguists, anthropologists, and community stakeholders, and it will integrate the insights from decades of linguistic inquiry.

Richard Futrell is Associate Professor of Language Science at the University of California, Irvine. His research integrates information theory, computational linguistics, and psycholinguistics to model how efficiency and memory shape language, earning best-paper honors at ACL and CogSci. He received a Ph.D. in Cognitive Science from MIT and an MA and BA in Linguistics from Stanford.

Kyle Mahowald is Assistant Professor of Linguistics at the University of Texas at Austin and works at the intersection of linguistics, cognitive science, and natural language processing. His papers have been recognized with Best and Outstanding Paper Awards at the computational linguistics conferences EMNLP and ACL, and he is the recipient of a National Science Foundation CAREER award. He received his M.Phil in General Linguistics from Oxford in 2011 and his Ph.D. in Cognitive Science from MIT in 2016; his postdoctoral training was at Stanford in the Natural Language Processing group.

Thus, we advocate a third view (joining fellow travelers in linguistics, cognitive science, and philosophy who have advocated for some form of this middle ground, e.g., Smolensky, 1988; Pater, 2019; Portelance & Jasbi, 2023; McGrath et al., 2024; Millière, 2024; Potts, 2025; Chesi, 2025): language models are not a complete theory of language – in fact, no one has such a theory – but they are hugely informative about language and its structure, learning, processing, and relationship with the larger structure of the mind. Language models have set off an intellectual explosion in cognitive science, machine learning, philosophy of mind, and other fields, in which longstanding ideas have been overturned, novel ideas are emerging, and disciplinary boundaries are dissolving. Linguistics has a chance to stand at the center of this huge intellectual ferment, and would be remiss to isolate itself intellectually on the basis that language models don't look like existing theory. Language science can contribute (and in fact already has contributed) to the development of language models, and language models can contribute (and in fact already have contributed) insights about language. Norm and Claudette have a lot to talk about.

2. Statistical models of language have outperformed expectations

2.1. A brief history of statistical language learning

The generative school of modern linguistics arose from an argument against the idea that the structure of language could be learned from the statistics of language data. "Chomsky perhaps most famously illustrated this by contrasting (1) Colorless green ideas sleep furiously" with "(2) Furiously sleep ideas green colorless." Even though neither sentence makes semantic sense, (1) is clearly grammatical in a way that (2) isn't. As Chomsky (1957) put it: "The notion 'grammatical in English' cannot be identified in any way with the notion 'high order of statistical approximation to English.' It is fair to assume that neither sentence (1) nor (2) (nor indeed any part of these sentences) has ever occurred in an English discourse. Hence, in any statistical model for grammaticalness, these sentences will be ruled out on identical grounds as equally 'remote' from English."

The effective conclusion from these arguments was that linguistic structure could only be characterized in terms of formal systems, based on rules or constraints and operating over structured arrays of symbols (Chomsky, 1965). Early on, the expectation was that such systems would form the basis for language technologies such as machine translation and question answering systems (Hays, 1960; Winograd, 1972; Hutchins, 1981). It was believed that these formal systems should be constructed by linguists since the task of learning them from data was hard or impossible. Yet despite myriad efforts to build machine translation systems, grammatical parsers, and other tools, linguistic competence remained elusive for machines. Symbolic approaches that sought to elucidate rules and structures often proved unable to capture all the exceptions and complexity that characterized natural language as it is used.

By the late 1980s and 1990s, statistical learning had a major renaissance in natural language processing (NLP) (Brown et al., 1990; Manning & Schütze, 1999; Pereira, 2000) and started to reappear in the human language learning literature as well (Saffran et al., 1996). Yet despite success at various NLP tasks, these techniques seemed frustratingly unable to get past approaches that simply counted up words and phrases (Wang & Manning, 2012; Arora et al., 2017). The connectionist movement in the 1980s and

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

90s seemed promising (Rumelhart & McClelland, 1986; Smolensky, 1988; Elman, 1990a) (and turned out to be the most important forerunner of modern language models), but there were well-justified concerns about the ability of these approaches to scale up and to represent the rules and structures that seem to characterize language (Pinker & Prince, 1988), despite accounts of how connectionist models could in principle implement rule-like symbolic behavior (Smolensky, 1990).

Even researchers optimistic about the role of statistical learning in language acquisition were, in the recent past, deeply skeptical that end-to-end neural approaches would succeed on interesting linguistic tasks. Representatively, Chater et al. (2006) wrote, "connectionism is no panacea here; indeed, connectionist simulations of language learning typically use small artificial languages, and, despite having considerable psychological interest, they often scale poorly." Tenenbaum et al. (2011) wrote, "connectionist models sidestep these challenges by denying that brains actually encode such rich knowledge, but this runs counter to the strong consensus in cognitive science and artificial intelligence that symbols and structures are essential for thought." Nevertheless, given the way that linguistics and machine learning developed and evolved, it now sometimes seems to be taken for granted in practice that ideas from linguistic theory will not form the basis of proficient language processing systems.

Throughout the statistical renaissance, generative linguists continued to claim that statistical methods would never solve interesting problems related to learning linguistic structure. In a review article, Everaert et al. (2015) reiterated the claim that statistical approaches would be fundamentally unable to distinguish sentences like "How many cars did they ask if the mechanics fixed?" from ungrammatical ones like "*How many mechanics did they ask if fixed the cars?," or that they would only be able to do so if provided with certain built-in formal structures. Berwick et al. (2011) were skeptical that recurrent neural networks (RNNs) could ever be much more powerful than bigram models.

Against this backdrop, the fortunes of connectionist models started changing in the 2010s – as a result of new techniques and, most importantly, due to increased computational power that made training neural models much more efficient (Hinton et al., 2006). By 2016, neural models showed rudiments of grammatical generalizations like subject–verb agreement (Linzen et al., 2016). Over the coming years, the success of models at acquiring linguistic abilities continued to grow (Futrell et al., 2019a; Wilcox et al., 2018; Manning et al., 2020; Hu et al., 2020; Warstadt & Bowman, 2022; Mahowald et al., 2024).

The growth from early neural models in, e.g., 2011 to now is remarkable from a historical perspective, and was surprising to virtually everyone in the field at the time. Sutskever et al. (2011) introduced an at-the-time state-of-the-art RNN that produced output like "In the show's agreement unanimously resurfaced. The wild pasteured with consistent street forests were incorporated by the 15th century BE" and praised it as a "surprisingly good language model," noting the "richness of their vocabularies," that the "text is mostly grammatical," and "parentheses are usually balanced over many characters." Compared to today's language models, the consistent street forests seem a long way away: 2011 might as well have been the 15th century BE. Today's language models (e.g., ChatGPT or Claude) produce seemingly endless streams of highly fluent and grammatical (if not always perfectly humanlike) text and form the basis of the most successful NLP approaches.

2.2. Neural LMs learn nontrivial linguistic structure

The rapid development of these capacities has been treated with healthy skepticism. LMs have access to vast amounts of data and could be simply memorizing or relying on cheap heuristics – giving the appearance of linguistic competence without really having it.

The most naïve way that one might think to test for grammatical competence in language models would be to see whether they assign higher probability to grammatical strings than ungrammatical ones. However, this would not work, even for an ideal statistical model, for reasons related to the point that Chomsky (1957) was making. For example, the ungrammatical string "Snails died the old" has a probability of $\sim 2^{-49}$ under the GPT-2 language model, whereas the grammatical string "The ancient crustaceans expired" has a lower probability of $\sim 2^{-55}$, simply because it uses lower-frequency words (Wilcox et al., 2023a).

But the failure of this naïve comparison does not mean that language models lack grammatical knowledge. In general – for language models and in real language use – many factors influence the probability of a string, of which the grammar of the language (as represented in the language model or in the mind of a human speaker) is only one: word frequency, utterance length, online processing constraints such as memory limitations, and plausibility given world knowledge all feed into the probability of an utterance. Indeed, not only utterance probabilities, but also comprehension accuracy, reaction time, and indeed any psychometric dependent variable are affected by all of these factors jointly – including the subjective grammaticality judgments that form the basis of much of the formal syntax literature (Kluender & Kutas, 1993; Hofmeister et al., 2013; Mahowald et al., 2016; Lau et al., 2017). Therefore, if we want to look for evidence that language models represent linguistic structure, we need to search more carefully.[1]

The key is to isolate linguistic structure from these other factors through controlled experimental studies and through probing LMs' internal states. Experimentally, from sufficient performance data, one may infer an underlying formal cognitive structure, no matter whether the implementation substrate is a brain or a neural network (Piantadosi & Gallistel, 2024). This is the standard procedure in linguistics, where data consisting primarily of acceptability judgments is used to postulate underlying linguistic competence. This approach can be applied just as well to LMs.

One form of such a study is to conduct behavioral comparisons of minimally different sentences. For example, LMs may be fed sentences such as "The keys to the cabinet are on the table" and the ungrammatical "The keys to the cabinet *is* on the table," and the conditional probabilities assigned to the verb form "are" versus "is" are compared. If the model understands how agreement works in English, then the probability for "are" should be higher than "is". This comparison controls for plausibility, since the context is the same for the two alternatives. The task could not be solved by any n-gram model with $n < 5$ (and higher-order n-gram models can be ruled out by adding more words between "keys" and "is/are" in the stimuli). The lexical frequency of "is" versus "are" can be controlled through a more elaborate experimental design, with four conditions in a 2×2 design crossing the grammatical number of the subject with the grammatical form of the verb (as done by Marvin & Linzen, 2018). These are usual procedures in psycholinguistics, where experimenters develop controlled sets of sentences and then measure dependent variables like reading times. The same methodology can be applied to language models

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

with probability as the dependent variable, aiming to ascertain whether models are sensitive to linguistic structure (Linzen et al., 2016; Futrell et al., 2019b).

Such studies have revealed behavioral patterns consistent with neural networks representing formal linguistic structures, in cases such as subject–verb agreement (Linzen et al., 2016; Bernardy & Lappin, 2017; Gulordava et al., 2018), filler–gap dependencies (Wilcox et al., 2018, 2023a; Kobzeva et al., 2023; Suijkerbuijk et al., 2023), and recursive embedding of clauses (Futrell et al., 2019b; Wilcox et al., 2019a; Hu et al., 2020), all of which involve highly nontrivial formal structures which statistical models failed to capture in previous work. Example results for subject–verb agreement from GPT-2 are shown in Figure 1: we see that grammatical verb forms are relatively more probable than matched ungrammatical verb forms in all but a few cases (human accuracy in producing the right verb forms in such sentences is 85%; Marvin & Linzen, 2018).[2] The results indicate that the model can represent the non-local structural dependency between the subject of the sentence and the matrix verb.

Of course, this work has not all been positive in its results. Models can be "right for the wrong reasons" (McCoy et al., 2019), adopting shallow heuristics which make correct predictions on certain test sets but do not generalize properly, or models may show mastery of linguistic form without a concomitant ability to understand the implications of utterances (Weissweiler et al., 2022) – in effect mastering linguistic form without mastering linguistic function (Mahowald et al., 2024). As an example of how shallow heuristics may lead to the illusion of good performance, a simple *n*-gram model performs fairly well on some subsets of paired grammatical and ungrammatical sentences of the BLiMP dataset (Warstadt et al., 2020) (although not as well as a neural LM), which does not explicitly control for *n*-gram frequency (Vázquez Martínez et al., 2023).[3]

The possibility that models generalize on the basis of shallow heuristics motivates deeper investigation. We discuss two approaches here. The first involves controlling a model's training data and then observing its generalizations on evaluation data that is unlike anything in the training data (Jumelet et al., 2021; Feng et al., 2024b; Misra & Mahowald, 2024; Leong & Linzen, 2023; Yao et al., 2025). This approach allows one to study true *generalization* in language models, rather than shallow memorized patterns. Ahuja et al. (2025) trained neural language models on a corpus of English-like text that has been constrained so that subjects and verbs are always adjacent. That is, the corpus contains sentences like "I saw the key to the cabinets," but never something like "The key to the cabinets is on the table". The question is, then, if the model trained in this way will generalize in the way that humans do, preferring "The key to the cabinets is on the table" (where the form of the verb depends on the subject of the sentence) over something like "The key to the cabinets *are* on the table" (where the form of the verb depends on the linearly previous verb, which is perfectly consistent with the training data). Ahuja et al. (2025) find that neural language models do make the human-like generalization, providing evidence that they can learn the formal structure underlying nonlocal subject–verb agreement without ever observing nonlocal subject–verb agreement (see also Patil et al., 2024).

Another approach is to dig into the language models' internal states. Large language models have a reputation of being black boxes, whose internal processes and representations are inscrutable, but researchers have been gaining growing traction on understanding their inner workings. Approaches include probing, where one attempts to decode linguistic features from the internal representations of neural networks, and causal interventions, where model internals are changed and the resulting output changes are observed (Hewitt & Manning, 2019; Chi et al., 2020; Voita & Titov, 2020; Manning et al., 2020; Papadimitriou et al., 2021; Ravfogel et al., 2021; Lampinen, 2024; Diego-Simón et al., 2024, among others). Such studies have revealed increasing evidence of grammatical structure in models. We consider these kinds of studies, particularly ones that probe models *causally*, to be promising avenues for developing linguistics and cognitive science (see Section 4.2).

A caveat to these findings is that the various evaluations of structural representation in language models have mostly used English (or a handful of other languages) as the target (Blasi et al., 2022). An important area for future work is to extend the empirical scope of these studies to a wider set of languages (work which is ongoing; see Jumelet et al., 2025, for a multilingual grammatical benchmark).

While there is disagreement about how much language models capture more complex formal patterns (Vázquez Martínez et al., 2023; Lan et al., 2024; Someya et al., 2024) or to what extent they can be said to "understand" (Bender & Koller, 2020) or refer to things in the world (Mandelkern & Linzen, 2024; Lederman & Mahowald, 2024), this is not the right place to adjudicate these debates. What is clear is that language models have learned nontrivial formal linguistic patterns better than many expected was possible. That is to say, they really *have* learned something about linguistic structure, in a way that problematizes earlier claims about the role of statistics in language learning. They are not *purely* relying on heuristics or memorization or cheap tricks. They have, to a meaningful extent, learned "the real thing" – that is, the thing that we care about, as language scientists interested in questions like how languages are learned, how they are processed, how and why they vary, and where they come from.

We next ask why this is interesting for the science of language and how the science of language productively integrates these insights.

3. The success of LMs is interesting for the science of language

> ". . . nor did Pnin, as a teacher, ever presume to approach the lofty halls of modern scientific linguistics, that ascetic fraternity of phonemes, that temple wherein young people are taught not the language itself, but the method of teaching others to teach that method; which method, like a waterfall splashing from rock to rock, ceases to be a medium of rational navigation but perhaps in some fabulous future may become instrumental in evolving esoteric dialects – Basic Basque and so forth—spoken only by certain elaborate machines."
>
> Vladimir Nabokov
> *Pnin*

LMs learn nontrivial aspects of language. But one could still object that this result is not relevant to language science because human language is the product of the human brain (and/or vocal tract, body, social environment, or what have you), whereas LMs have a totally different architecture and different objective. Perhaps these differences are so great that a language scientist has practically nothing to gain from studying language models, in the same way that there is limited value to an ornithologist studying jet engines, even though both bird wings and jet engines enable flight (Kodner et al., 2023).

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

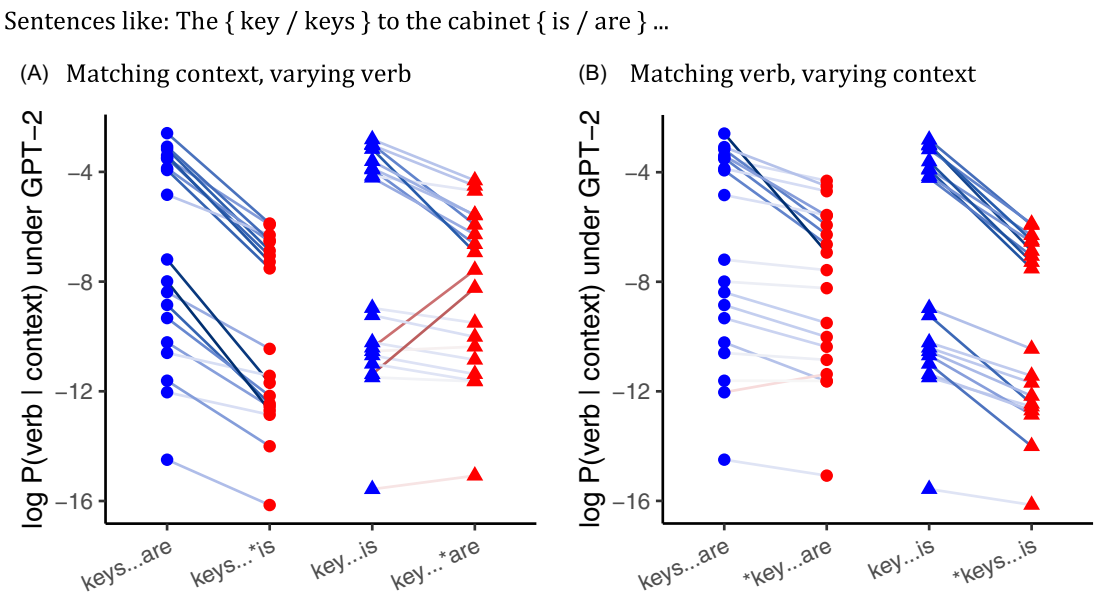


Figure 1. Language models show sensitivity to linguistic structure. Figure shows data from targeted evaluation of subject–verb agreement in GPT-2 from Gauthier et al. (2020), based on design and materials from Marvin and Linzen (2018). Each point represents the log conditional probability for a verb (e.g., singular *is* or plural *are*) in a particular sentence. **A**. Blue points are for grammatical verbs, and red for ungrammatical, in matching contexts. The difference in log probability between grammatical and ungrammatical verbs in identical contexts is indicated by a line. For plural subjects, the grammatical verb is always higher probability. For singular subjects, the grammatical verb is higher probability in 79% of cases. Human accuracy on this task is 85% (Marvin & Linzen, 2018). **B**. Blue points for grammatical verbs and red points for the same verbs in a context that makes them ungrammatical. The correct context makes the grammatical verb higher probability in all cases for the singular verb *is*, and in 95% of cases for the plural verb *are*.

Here we argue that the success of language models is relevant, indeed important, to the science of language because (1) there are good reasons to believe that there should be parallels between engineering artifacts like language models and human language, (2) the success of language models in learning from text upends ways of thinking that are deeply ingrained in generative linguistics and parts of cognitive science, and (3) language models are already connected to statistical and probabilistic traditions in linguistics, and in fact arose out of them to a large extent. As such, LMs aren't alien invaders into linguistics from engineering. Rather, they are tools similar to those that have long been used to answer fundamentally linguistic questions.

3.1. Parallels between engineering models and cognition

From the study of vision, there is strong precedent for the idea that neural networks developed purely for practical applications can tell us a great deal about cognition as it is implemented in the brain. Similarly, there is good reason to think that LMs will be informative for the study of human language. To elucidate this example, we briefly review, in simplified form, the development of the neuroscience of vision and the role of neural networks in it.

Hubel and Wiesel (1959) discovered that early processing of visual information is performed by neurons that are selectively responsive to *edges* in the visual input. After this discovery, the question became why early visual processing works this way. Olshausen and Field (1996), building on information-theory-inspired intuitions from Barlow (1961, 1989), showed that edge detectors resembling those found in the visual cortex were the generic solution to the problem of representing visual information accurately in a neural network under a constraint that only a small number of units should be active (that is, a sparsity constraint, arising ultimately from a power constraint on neural firing). Even better explanation – both in the sense of explanatory depth and predictive accuracy – came from the engineering of artificial visual systems, in particular the development of AlexNet, a large (for the time) hierarchical convolutional neural network for image classification (Krizhevsky et al., 2012). Yamins et al. (2014) showed that this architecture, when trained to do object recognition, not only developed edge filters in its early layers but also receptive fields in later layers corresponding to later layers of visual processing in the primate brain. The overall picture that emerges from this line of work is: the neuronal organization of visual processing is determined by the function of finding a sparse code to identify and manipulate objects in the environment, a function largely shared between biological and artificial systems.

Here, the artificial system provided guiding insights for understanding the natural system.[4] Something similar is now happening in the science of language processing, where the internal representations developed by language models are predictive of activation patterns in language areas of the brain (Goldstein et al., 2022; Caucheteux et al., 2023; Hosseini et al., 2024b; Rathi et al., 2024), and predictability as estimated by a language model is an important factor in studies that predict neural activity (Stanojevíc et al., 2023; Zhao et al., 2025). The picture is not yet as clear as it is in vision – this is an area of active research, and there are no animal models we can use to get the plentiful high-resolution controlled neural data we would like – but there is precedent to think that the artificial system will be informative about the natural system here too.

Why did this work in vision? To explain the success of the approach, Cao and Yamins (2021) introduce the *Contravariance Principle*: the idea that, if we want to uncover solutions to problems that are common between brains and models, we should focus on hard problems. The reason is that, if a computational problem is

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

hard in the sense that it requires satsifying multiple potentially competing constraints at once, then there are likely to be only a small number of ways to solve it. So we expect different systems that solve the same hard problem (e.g., neural networks and the brain) to converge to the same solution (see also Huh et al., 2024; Hosseini et al., 2024a). If a problem is relatively simple, then we might expect many different solutions to work. For instance, there are famously scores of algorithms that have been proposed for sorting numbers in an array.

Vision is not like sorting a list of numbers: it is a hard problem that requires satisfying many constraints (e.g., fast processing, reliable transmission of input, invariance to different light conditions, among many others). Therefore, we can expect that *any* system that solves the problem will share core features with any other. We see this not only in the comparison of humans vs. neural networks but also in the comparison of humans vs. monkeys (Rajalingham et al., 2018) and primates vs. animals whose visual cortices arise from totally different evolutionary phylogenies, such as cephalopods (Pungor et al., 2023).

The problem of learning and using language is also not like sorting a list of numbers: an entity that learns and processes language has to satisfy many constraints (storage of lexical items, generalization to novel contexts, fast processing, etc.). And as we have discussed, large transformer-based models do learn and deploy nontrivial aspects of language. If the Contravariance Principle is right, then it is reasonable to expect some degree of commonality in the solutions that exist in the brain and in transformers. Indeed, the massive literature on linguistic interpretability in transformers, which we partially reviewed in Section 2.2, has revealed representational strategies in neural networks that lead to correct novel predictions about human performance (Lakretz et al., 2021), and the way that transformers process syntactic features like agreement cues has close parallels with independently proposed cognitive frameworks for modeling human language processing based on cue-based retrieval (Lewis & Vasishth, 2005; Ryu & Lewis, 2021; Timkey & Linzen, 2023).

Thus, it is just not the case that these models can be dismissed out of hand because they are different from the brain. To do so would be to miss out on a source of *in silico* insight and inspiration which has been hugely useful in other areas of cognition.

3.2. Understanding LM success requires rethinking language learning

3.2.1. The significance of the learning problem in linguistics and cognitive science

For a century, cognitive scientists, linguists, computer scientists, and philosophers developed formal theories of how language learning – or any learning – must work. The key concept has been what knowledge is brought to the learning process by a learner beyond the data, known as inductive bias in the machine learning literature (Mitchell, 1980; Goyal & Bengio, 2022), as illustrated in Figure 2. We can think of a learner as a device that takes in some data and outputs a hypothesis, or a set of hypotheses, or a probability distribution over hypotheses, for the underlying process generating the data. For example, given a bunch of data points in a 2D plane, one hypothesis would be that the points are generated along a line in the plane; this is the assumption underlying linear regression. Given linguistic utterances as data, one might hypothesize a certain grammar underlying the data.

The key issue is that, in a very general sense, the data fundamentally underdetermine the hypothesis that the learner arrives at, because any data is compatible with many, many hypotheses. In Figure 2A, where the learning data is points and the hypotheses are lines, both the straight and the curvy lines fit the learning data equally well; the fact that the straight line seems like the right hypothesis – and is likely the one that humans would seize on – therefore cannot be a function of fit to the data. It must be a function of fit to the data plus something that biases a learner to favor one hypothesis over another, even among hypotheses that fit the data equally well: this is inductive bias. Inductive bias is something that the model (or modeler) brings to the problem, not something inherent in the data.[5]

Inductive bias shows up in several forms and under many names in the scientific literature. In linguistic theory, it appears perhaps first as the *evaluation procedure* in Chomsky (1965, pp. 30–48), a function which compares two different grammars that can generate an observed set of sentences (that is, two grammars that fit some learning data equally well), and ranks them in order of preference, in a way that captures how humans generalize beyond the data (Chomsky, 1965, p. 45). More generally, it appears as Universal Grammar (UG) – an "innate schematism of mind that is applied to the data of experience" (Chomsky, 1971, p. 28) that enables language learning and generalization; furthermore, UG is held to be domain specific to language (not applying to any other aspect of cognition) and species specific to humans (Huybregts, 2019). In this approach, (generative) grammatical theories such as Minimalism are hypotheses about the nature of UG (Chomsky, 1993; Adger, 2003). They are meant to precisely delimit what languages may exist and be acquired by humans.

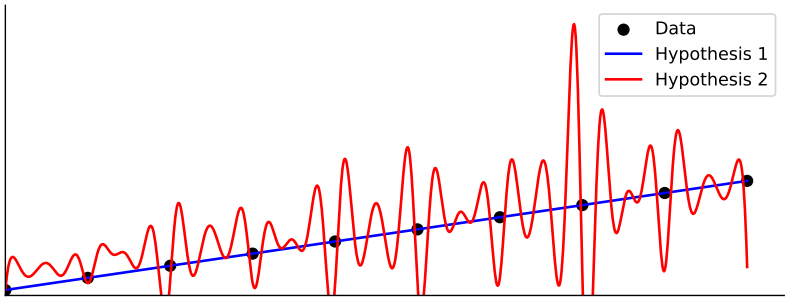
This approach to linguistic theory serves the goal of *explanatory adequacy*: the idea is that a theory of grammar should not only capture which sentences are grammatical and ungrammatical in a particular language, but also that the theory encompass all and only the possible languages that we might actually find in the world (Chomsky, 1965, Ch. 1). Thus, if a grammatical theory seems to accommodate "impossible" or unattested kinds of languages, then this would be considered evidence against that theory, because the theory is supposed to represent the universal and language-specific "schematism of mind" that a learner applies to language data.[6] This approach to explaining linguistic phenomena has been called "explanation by constrained description" (Haspelmath, 2009, pp. 384–385).

Naturally then, UG has been the major focus of study in language learning, seen as the simultaneous solution to two different problems: (1) how children can learn language from inadequate data (the classic Argument from the Poverty of the Stimulus: see Pearl, 2022, for a review) and (2) why human language is the way it is: because UG strongly restricts the set of possible languages.[7] The dream has been to come up with a formalism for linguistic description which captures human generalizations, is domain-specific to language, and arises from a genetic endowment unique to humans (Chomsky, 1988; Hauser et al., 2002; Berwick et al., 2011) – thus in one fell swoop solving (1) and (2).

Under this logic, it follows that a good theory of language learning should be *restrictive*: that is, there should be languages that *cannot* be learned under the theory, and this restriction on the hypothesis space provides explanatory adequacy.[8] In terms of Figure 2, the intuitively bad hypotheses would be ruled out as simply unavailable as mental representations during learning

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

(A) Numerical data



(B) Linguistic data

Data	
The key is pretty.	The keys are pretty.
Are the keys pretty?	I like the keys to the cabinet.
The man is happy.	Is the man happy?
The man who is tall is happy.	I like the man who is tall.

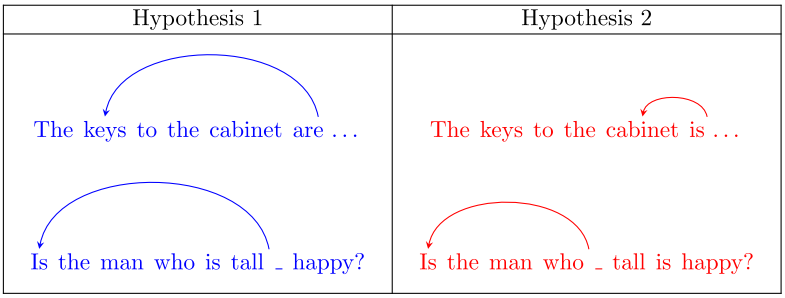


Figure 2. Illustration of the role of inductive bias in generalization. **A.** The black points represent data, and the lines represent generalizations that could be formed on the basis of the data. Both lines fit the data exactly, but make different predictions for new datapoints. Intuitively, the blue line seems like a better hypothesis. But the choice to prefer any single generalization over any other can only be a function of *inductive bias* – a preference for hypotheses which are in some sense "simpler" – which is not a function of the data. **B.** Inductive bias in language. In black, possible corpus data from English. In blue, sentences which generalize from the corpus data on the basis of hierarchical structure. In red, sentences which generalize on the basis of linear word order. Intuitively, the blue sentences seem like natural generalizations, and the red ones seem unnatural, even though all are equally consistent with the data given. This phenomenon is claimed to be the result of innate biases specific to language (Chomsky, 1971).

(Everaert et al., 2015).[9] This approach to linguistic explanation has a pleasing elegance to it: learners *must* be restricted to learn properly, and we see that the variation in actual languages is restricted, therefore we can kill two birds with one stone by finding the right representational restrictions on learners.

3.2.2. The modern view on learning

The logic of inductive bias is sound. On a deep level, there really is no free lunch in learning, even deep learning (Mitchell, 1980; Wolpert et al., 1995; Baxter, 2000; Adam et al., 2019). Asking how much one can learn "from the data alone" without inductive bias is like asking how close a runner can get to the finish line without saying where they started. It's just not a sensible question. However, developments in deep learning have forced revisions to conventional ways of thinking about how inductive bias arises.

The conventional wisdom was that a learning model must be restricted in terms of the hypotheses it can consider, so that it does not overfit (Bishop, 2006, §1.1). A relatively unrestricted model may memorize the training data, or find some other pathological solution which lurks in the depths of its vast hypothesis space, as in the red hypotheses in Figure 2. By restricting the hypothesis space, one provides strong inductive bias, so that pathological solutions can be ruled out and good generalization can be guaranteed. The logic is the same as UG: explaining generalization requires that we restrict the set of hypotheses (grammars) that a model (learner) may entertain.

However, the empirical finding from deep learning is that *overparameterized* networks, which are more than flexible enough to memorize their training data, do generalize, often better than relatively restricted ones (Belkin et al., 2019; Zhang et al., 2021). In machine learning, overfitting has typically been measured in terms of performance on a held-out dataset, where one typically sees a U-shape curve as in the left part of Figure 3: as models become more flexible in terms of the hypotheses they can express, at first they generalize to the held-out data poorly (underfitting), then they generalize well (the sweet spot), then after some point the more powerful models learn poorer hypotheses (overfitting). And yet in the late 2010s, it was discovered that if you keep adding model power, the U-shape curve starts to descend again, in a phenomenon called *double descent*, shown in Figure 3. This pattern suggests that learners continue to find good generalizations even *after* being flexible enough to memorize their training data.[10]

It seems that what appeared to be an insurmountable barrier turned out to be a kind of phase transition in how learning works.

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

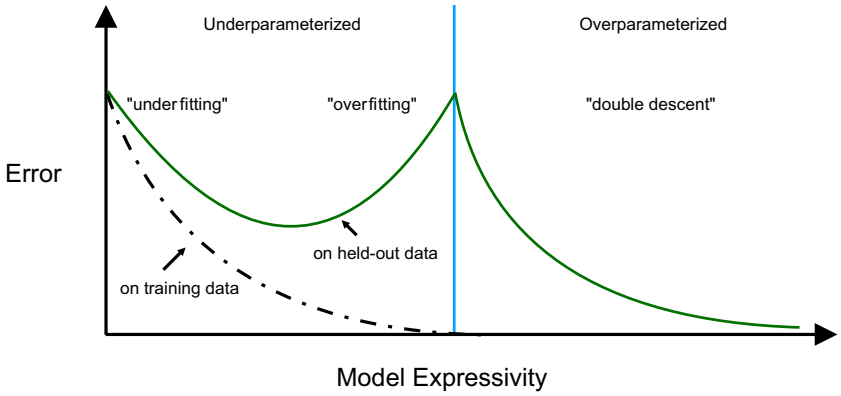


Figure 3. Expected error of a learning model on training set and test set as a function of model expressivity (the variety of hypotheses that the model can express). To the left of the blue line, we have the classical picture from statistical learning theory. In the *underfitting* regime, increasing model capacity reduces training and test error, up to a point where one enters the *overfitting* regime, in which increasing model capacity causes a decrease in error on the training data but an increase in error on held-out data. This overfitting phenomenon is intuitively due to the model gaining the capacity to memorize the training set without forming generalizations that would be useful on the test set. The goal of model fitting in this view is to find the *sweet spot* that minimizes test error, often accomplished through methods such as deliberately reducing model capacity or early stopping. However, modern deep learning has revealed that as model capacity increases *beyond* the point where the model can memorize the training data (the *interpolation threshold*, in blue), we enter a new *overparameterized* regime, where test error decreases, due to soft simplicity biases in the learner (Maddox et al., 2020). Figure based on Belkin et al. (2019).

What looks like overfitting is, counterintuitively, ameliorated by adding even more parameters. Thus, one sees a new kind of generalization emerging, where less restricted learners have superior performance in learning the right generalizations.

These findings came as a surprise to statistical learning theorists (Belkin et al., 2019; Zhang et al., 2021) and have triggered an ongoing field-wide effort to rethink learning theory or to show how these unexpected findings are compatible with existing theory (for example, Poggio et al., 2020a; Martin et al., 2021; Martin & Mahoney, 2021; Kuzborskij et al., 2021; Henighan et al., 2023; Attias et al., 2024). These results should also make us rethink language learning and the role of restrictive formalisms in linguistic explanation. It is simply not the case that proper generalization can only come from learners who are sharply restricted to small hypothesis spaces, nor even that there is a correlation between restrictedness and proper generalization. All the successes of modern machine learning are in contradiction to the view that generalization must come from hard restrictions on the hypothesis space.

But how can this be, if the idea of inductive bias is right? How did the conventional logic go wrong in practice? The key mistake was the conflation of model power with inductive bias (Hubinger, 2019). The modern empirical finding is that *more* flexible learners have *stronger* biases toward "simple" hypotheses (Goldblum et al., 2024): that is, although they do not restrict the range of hypotheses that *can* be considered, they *do* impose a soft notion of simplicity on those hypotheses. The logic is illustrated in Figure 4, following Wilson (2025). Furthermore, the source of this simplicity bias in neural networks and related systems is not necessarily related to the hard limits of their expressivity. The simplicity bias in neural networks seems to be toward functions that are nearly linear or simple in other ways (Valle-Perez et al., 2018; Hahn et al., 2021b), likely as a result of the dynamics of gradient descent on the loss landscape induced by the model (Poggio et al., 2020b; Pezeshki et al., 2020; Merrill et al., 2021; Hahn & Rofin, 2024).

In machine learning, these discoveries have catalyzed a change in focus, away from models whose architecture and representations are tailor-fit to their domain, and toward models that learn quickly in relatively unrestricted hypothesis spaces (Sutton, 2019). The zeitgeist is now strongly against using domain knowledge to restrict the behavior of learners, even when researchers have a strong sense of what the relevant domain knowledge is.

3.2.3. The upshot for linguistic theory

These surprising results from deep learning mean that the research program of achieving explanatory adequacy through restricted learners is deprived of its apparent logical inevitability. Previously, the logic was: learners *must* be formally restricted, and we see that human languages *do* have universal patterns, so maybe these necessary restrictions create the universal patterns. Nowadays, if this approach to linguistic explanation is to be maintained, the logic must be: learners have hard formal restrictions *even though this is not necessary and may be harmful for learning*, and these restrictions create language universals. It is still a viable hypothesis, but it loses its elegance.

Furthermore, even if language-specific innate inductive biases in humans really are the key to the structure of human language, these inductive biases might not be expressible in terms of a categorical symbolic formalism or a sharply limited hypothesis space for learners. Inductive biases in modern neural models are soft and seem to arise from a complex interplay of training dynamics, objective function, and model architecture, with the hard limits of model expressivity playing a relatively minor role.

Overall, "explanation by constrained description" no longer seems so explanatory, at least not without further stipulations. Language learning and language universals may well be better captured by a highly flexible, less constrained formalism for linguistic representation – one which on its own could capture non-linguistic patterns as well as natural linguistic ones – paired with a soft, quantitative simplicity metric that captures learning dynamics or functional pressures on language. Learning models with explicit simplicity biases exemplify this approach (e.g., Hsu et al., 2011; Perfors et al., 2013; Rasin et al., 2021; Lan et al., 2022).

This is not to say that inductive biases are no longer important in linguistics and language learning, nor that investigating inductive biases of humans and neural networks is not important: we will have much more to say on this in Section 4.4. Far from demoting inductive bias as a concern, language models open up the range of (possibly innate) inductive biases that we might look for in humans. The main point for linguistic theory is not to demote the importance of the learning problem, but to reshape it: away from explanation by constrained description, and toward a broader landscape of approaches and hypotheses.

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

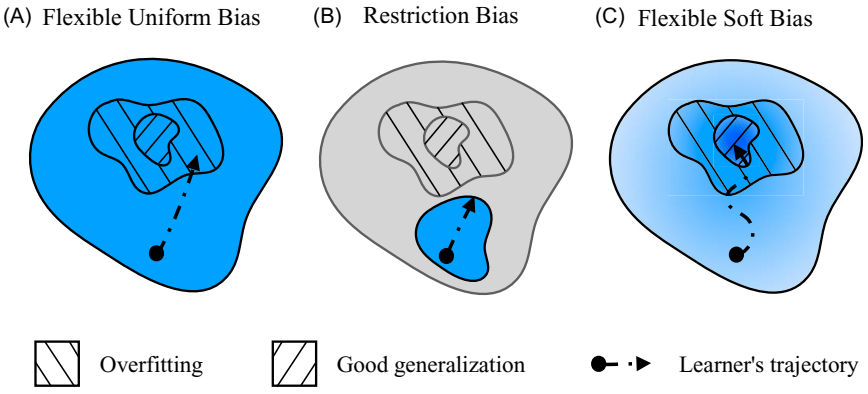


Figure 4. Illustration of the logic of a soft inductive bias within a flexible hypothesis space, adapted from Wilson (2025). Figures show a hypothesis space for a learner, including regions of bad generalization (overfitting) and good generalization. The learner starts at the dot and, throughout learning, moves in the direction indicated by the arrow. **A.** Given a uniform bias over a highly flexible, relatively unrestricted hypothesis space, the learner is likely to overfit and form pathological generalizations. **B.** One strategy to impart useful inductive bias to a learner is to restrict the hypothesis space. However, this may miss the good generalizations. **C.** A better strategy, which is part of what underlies neural network success, is to keep a highly flexible, relatively unrestricted hypothesis space, but impart a soft simplicity bias which guides the learner towards the good generalizations.

3.2.4. The question of data quantity

It is still the case that, whatever language models learn about language, they do it using orders of magnitude more linguistic input data than human children are exposed to (Yedetore et al., 2023; Warstadt et al., 2023). Moreover, the learning trajectories of models and humans continue to show systematic differences (Chang & Bergen, 2022; Evanson et al., 2023; Constantinescu et al., 2025). Taken together, these results suggest meaningful differences between learning in models and humans, both in terms of data requirements and patterns of learning. A form of the Poverty of the Stimulus argument is still alive in the form of the observation that even if neural networks acquire linguistic structure, they do not do so on the basis of the same amount and kind of data that a human child learner gets (Lan et al., 2024).

As reviewed above, it has been suggested that formal linguistic structure provides the key inductive bias that makes learning from small amounts of data possible. However, there is no evidence that linguistic structure is the major missing ingredient to close the gap between models and human children.[11] There have been many attempts to inject linguistic structure into neural models in various ways; some of the earliest successful deep learning approaches to language understanding were based on architectures that performed recursive computation in hierarchical parse trees (Socher et al., 2011, 2013). These approaches do succeed in promoting the kinds of structural generalizations that humans find intuitive, with the result that language learning from data is somewhat more sample efficient (Dyer et al., 2016; Futrell et al., 2019b; Wilcox et al., 2019b; Kim et al., 2019; Papadimitriou & Jurafsky, 2020, 2023; Nandi et al., 2025). However, in terms of absolute performance, these approaches are still far from closing the gap between humans and neural networks.

It is entirely possible that the key to faster and more human-like learning is not particular built-in formal constraints, but rather *more flexible* architectures (such as Kolmogorov–Arnold Networks: Liu et al., 2024) or biases toward domain-general compositional reasoning (McCoy & Griffiths, 2023; Yang & Piantadosi, 2022) or different training regimes (Murty et al., 2023) or the incorporation of multimodal data which provides rich side information about the structure of the environment that is being described in language (Wang et al., 2023) or domain-general bounded-rational approaches to generalization such as the Tolerance Principle (Belth et al., 2021; Payne et al., 2021; Kodner, 2022).

Overall, we believe the shoe is on the other foot for those who believe that an innate bias toward a particular formal linguistic structure is the factor that enables human learning from small amounts of data. Formal linguistics has not presented an alternative model with the demonstrated practical language-learning capabilities of neural models. The direction of developments in machine learning suggests that the gap between human and machine learning is more likely to be closed not through built-in domain-specific formal restrictions, but rather through more powerful domain-general learning algorithms.

The difference between human and neural network learning also raises the possibility that humans and neural networks are just so different that one is not informative about the other (Kodner et al., 2023). We believe not. First, as argued above, the success of neural networks weakens *logical* arguments that language cannot be learned without domain-specific formal constraints on language, and in general changes how we think about the role of formal restrictions in learning. Second, *even if* standard neural network training methods are not able to acquire linguistic structure on the basis of developmentally realistic data, the representations that they do eventually acquire based on more data are still informative about how language might be represented and processed in the brain (more in Section 4), even if the networks do not arrive at these representations along the same trajectory as the human child.

3.3. Language models and linguistic traditions

While language models may seem to involve concepts that are foreign to linguistics, they emerged in part from certain intellectual traditions of the study of language – although these traditions are not necessarily the ones that have been dominant in linguistics departments. Below we discuss the relationship between schools of linguistics and the development of language models and related language technologies.

3.3.1. The generative tradition of linguistics

It is rare in the history of science for a scientific theory to turn out as disconnected from a corresponding engineering application as formal generative linguistics has turned out to be for language models. We believe this has happened primarily because of a difference in goals between generative linguistics and language modeling, with the generative tradition taking a narrow view of its goals.

A historical parallel is informative. In the early 19th century, while Newtonian mechanics provided a strong basis for parts of mechanical engineering, it did not on its own seem to provide a theory that answered pressing questions about the increasingly complex machines emerging from the Industrial Revolution, in particular steam engines. Carnot (1824) developed an effective theory of such engines using a new ad hoc concept of "moment of

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

activity," which eventually developed into the idea of entropy (Clausius, 1865). This concept had to be discovered through engineering because the focus of purely theoretical physics was on understanding fundamental mechanics through mathematical methods of increasing elegance. Explaining the behavior of complex machines – which behavior certainly must reduce to more basic mechanics anyway – was not on the agenda. And yet from the practical project of understanding steam engines emerged a family of concepts that some theoretical physicists now view as more fundamental than even matter itself (Wheeler, 1989).

Turning back to linguistics: the reason that generative linguistic theory did not turn out to be helpful along the path toward language models was due to a similar narrowness in focus, which led it away from considering complex systems for dealing with language.

Language models are, at bottom, models of the stream of language that is produced and comprehended by humans. The relevant theory is the theory of language use, production, comprehension, learning, and of human cognition more generally. Generative linguistics, on the other hand, has not focused on how to build a machine that produces and comprehends language, but rather *how to build a language*, conceived of as an abstract mental structure that gives rise to a mapping between meaning (that is, a logical form or conceptual–intensional representation) and form (that is, a phonological form or sensorimotor representation) (Chomsky, 1995, 2005; Adger, 2003; Hornstein et al., 2005).[12] Furthermore, it was claimed that this kind of analysis must take center stage in the science of language, preceding any analysis of more complex systems for language processing, use, or learning, since these systems must operate in ways that make reference to the abstract structures of language (Chomsky, 1965, Ch. 1).

Given this focus, generative linguistics *has* had a huge influence on the engineering field that works on how to build languages: programming language design. Indeed, the basics of all modern programming languages are based on formal linguistic theory, incorporating explicit (often contextfree) grammars and parsers, as a way of linking a stream of text to a representation of a computation. Programming languages are, perhaps, exactly the "esoteric dialects" spoken by "elaborate machines" in the Nabokov quote that starts this section.

While this narrowness of focus was useful on its own terms, it has limits. There are potentially fundamental insights to be gained now from the analysis of messy, complex, practical systems, just as happened in physics, and as is now happening in linguistics.

3.3.2. The statistical tradition of linguistics

In fact, there is a long tradition of statistical approaches to language, and these traditions were deeply involved in the early development of language models.

Perhaps the most well-known such contribution is *distributional semantics*: the idea that the semantics of a word is related to the distribution over contexts in which that word appears, often cited to Firth (1957, p. 11), and developed more systematically by Harris (1954). This idea originates from the school of structuralist linguistics (Saussure, 1916; Bloomfield, 1926). Structuralist linguistics had as one theoretical aim the development of *discovery procedures*, which were formal procedures that could be applied to bodies of text in order to discover (and even *define*) linguistic structures (Harris, 1951). The most well-developed of these discovery procedures were statistical in nature. For example, Harris (1955) developed a theory of words and morphemes based on statistical co-occurrence patterns, including a procedure for discovering morpheme boundaries by effectively calculating transitional probabilities, an idea taken up again much later in the psycholinguistics of language learning (Saffran et al., 1996), and closely related to tokenization methods such as byte-pair encoding (Shibata et al., 1999; Tanaka-Ishii, 2021, Ch. 11). At issue was the relationship between grammatical structure and the observable statistical structure of a corpus of language – one of the core questions that linguists working on language models are interested in again today.

Harris's work is an intellectual precursor to modern language models. However, in practice, the statistical structuralist tradition represented by this work was largely supplanted in American linguistics departments by the generative school, which sprang from Chomsky's (1957) arguments that distributional statistics were irrelevant to linguistic structure and that discovery procedures were a distraction from the putatively "core" questions of linguistics.

Nevertheless, the statistical analysis of language continued under the heading of usage-based approaches (Bybee & Hopper, 2001), in which the key to language is not a set of underlying formal rules but emergent properties based on the statistics and dynamics of language use (from which rules or rule-like behavior might emerge). Traditions in typological syntax (Greenberg, 1963; Dryer, 1992), functionalist syntax (Comrie, 1989; Keenan & Comrie, 1977), construction grammar (Croft, 2001; Goldberg, 2009), probabilistic modeling (Bresnan et al., 2001, 2007; Christiansen & Chater, 2016), and evolutionary linguistics (Kirby & Hurford, 2002) have all carried the banner of a thoroughgoingly statistical approach to language. Earlier arguments about the relationship between statistics and structure and the value of probabilistic methods mirror many of the points we are making here (Abney, 1996; Pereira, 2000; Manning, 2003; Bresnan et al., 2007; Lappin & Shieber, 2007; Norvig, 2012).

These statistical traditions of linguistics played a catalyzing role in the birth of neural networks and LMs. The connectionist framework from which neural LMs emerged overlapped with the statistical tradition in linguistics, often in conflict with the anti-statistical tradition. A major locus of this early work using neural networks was about the English past tense, with Rumelhart and McClelland (1987) developing a neural model for forming the past tense from present tense words, triggering much debate (Pinker & Prince, 1988). Elman, a Linguistics PhD, developed an early RNN sequence architecture (Elman, 1990a) in order to solve the problem of finding linguistic structure in time, a problem motivated by findings in linguistics and psycholinguistics (Frazier & Fodor, 1978). This basic architecture still underlies many language models (Feng et al., 2024a). Chris Manning, a pioneer in neural and probabilistic models, was trained as a linguist (Manning, 1995) and argued for a statistical approach to syntax (Manning, 2003), before going on to work on influential neural methods for performing linguistic tasks (e.g., Pennington et al., 2014).

The history of the field and the dominance of the generative tradition made it such that these figures are sometimes seen as "not linguists". But the traditions they emerged from were fundamentally concerned with questions about human language and cognition and belong under the heading of linguistics. To be clear, we do not think that the success of modern language models *unequivocally* supports or refutes any particular intellectual tradition in linguistics. But we do think that the narrow and exclusive theoretical focus of generative linguistics, coupled with its relative dominance within American linguistics departments throughout the late 20th century, caused tragic missed connections

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

in intellectual history and has left the field of linguistics diminished compared to where it could be.

Progress in language science will not come from hyperfocus on any one goal, nor from any one theory of what language is, and nor from one framework for understanding linguistic phenomena. Rather, the science of language needs to draw on a plurality of perspectives, from a wide variety of disciplines. In the current moment, this means leveraging the explosion of intellectual creativity springing forth around language models. Just as the concept of entropy arose from ad hoc analysis of complex machines and ended up revolutionizing fundamental physics, it is possible and even likely that ideas based on language models will revolutionize linguistics.

4. Where does that leave the science of language?

> "Trying to understand perception by studying only neurons is like trying to understand bird flight by studying only feathers: It just cannot be done. In order to understand bird flight, we have to understand aerodynamics; only then do the structure of feathers and the different shapes of birds' wings make sense."
>
> Marr (1982, p. 27)

If neural models make us rethink some aspects of linguistic theory, where do we go from here? What is the status of the formal structures discovered through linguistic analysis, if they don't have to be innately latent in the human brain? We believe that linguistic structure is as real as it ever was and that the LM revolution presents us with important new ideas and methods for studying language. Taking a statistical and information-theoretic approach to language brings these possibilities into focus. We also consider ways in which LMs might give us reason to update our linguistic theories to be more gradient, usage-based, and functionalist, by serving as a proof-of-concept system that implements ideas that were previously hard to formalize.

4.1. Linguistic structure is real

LMs are proof-of-concept that systems can process language without having linguistic structure hardwired in. But that doesn't mean it isn't real. We hold that linguistic structure is a *real pattern* in the sense of Dennett (1989): it provides a compressed and useful representation of important aspects of language. To explain or describe linguistic phenomena without reference to linguistic structure, perhaps in terms of some more reductive neural theory, would be hopelessly complex – and would make understanding LM behavior more difficult (see Nefdt, 2023, for a worked-out theory of linguistic structure as real patterns).

Figure 5 shows a schematic of how this account works for a phenomenon like subject–verb agreement. Modern language models are proficient at this task, proficient enough that their behavior looks rule-like. But, given the dynamic and chaotic nature of neural networks, it is likely that there are messy and heterogeneous structures and processes that give rise to this rule-like behavior in models. These structures might only very noisily map onto linguistic categories like "grammatical subject" or "number agreement". One reaction to that might be to conclude that linguistic theory was wrong or misguided – that it seemed like there was such a thing as "grammatical subject" that was crucial for verb agreement, but that actually the underlying reality was more complicated and now we can do away with all these rules and replace them with the messy, complicated neural system instead.

This hyper-reductive conclusion would not be productive. A theory of language which points to complex neural circuits for explaining the data of subject–verb agreement, without referring to something about the concept of grammatical subject, would be like an example from Dennett (1989, Ch. 2): imagine aliens who can perfectly predict human behavior using atomic physics (that clump of human atoms will move toward that coffee shop) but without understanding human beliefs and desires (that the human wants coffee and believes that the coffee shop will provide it). Lacking concepts of humans' attitudes, the aliens' theory of human behavior is perhaps perfectly accurate and reductive, but comes "at enormous cost of cumbersomeness, lack of generality, and unwanted detail" (Dennett, 1990, p. 189). Dennett concludes that the concept of belief is a *real pattern* in that it provides an abstraction that supports prediction and counterfactual reasoning based on coarse-grained data. The idea is that any abstraction that enables simple prediction, compression, and causal modeling in this way is a real pattern, fully deserving the epithet *real* even if there is a more reductive theory that lacks the abstraction. Similarly, even if it turned out we could explain linguistic behavior without recourse to anything recognizable as linguistic structure, it would come at the same cost of "cumbersomeness, lack of generality, and unwanted detail".

We join Dennett in his conviction that real patterns are, well, real. As such, we are not eliminativists about grammatical structure, nor do we believe that this kind of eliminativism would be scientifically productive. Some in the generative tradition adopt a similar view, contra the idea that every posited linguistic structure must be in the head. Adger (2022) writes: "Generative grammarians think of the posits of their theories in much the same way as physicists think of quarks or particles in wave form: They are the best explanations of the phenomena, though we do not know exactly what they correspond to in the world we can currently observe." LMs are a promising avenue for figuring out what these explanations do, in fact, correspond to in the world.

Indeed, the idea of real patterns can help explain the success of language models. Language models are trained to predict tokens in context. From an information-theoretic perspective, this means they are finding simple representations for linguistic input (Chater & Vitányi, 2007), and in particular, compressed representations of past input that let them predict future input (Tishby et al., 1999; Shalizi & Crutchfield, 2001; Bialek et al., 2006; Still, 2014; Tishby & Zaslavsky, 2015). Insofar as linguistic structure is real, it is part of the representation that enables this compression.

A similar point is made by Wolfram (2023), who connects the success of LMs with the reducibility of computational processes. Some computational processes are chaotic: there is no good way to systematically predict the digits of pi other than actually running the computation and learning what the next digit is. But other processes that *appear* complex have learnable patterns that can be used to give simpler, compressed representations. For instance, $\frac{3227}{555}$ is 5.8144144144 . . . where the 144 repeats forever. This repeating structure means that we could give a compressed representation of what the millionth digit is without having to do additional computation. Wolfram suggests that the success of LMs proves that human language has structure that enables this kind of compression. This compressibility is responsible not only for LMs' success, but for the fact that it is possible to construct linguistic theories. Compressibility *is* structure.

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

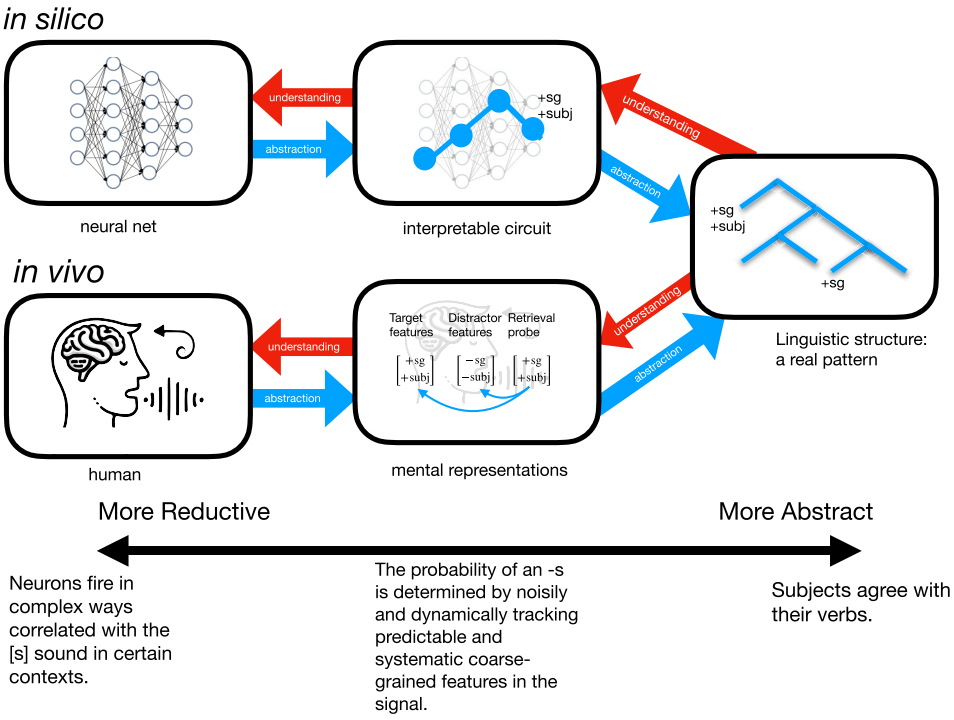


Figure 5. Schematic for linguistic structure as a *real pattern*, a (potentially leaky) abstraction that supports prediction and counterfactual reasoning based on coarse-grained data. In humans (*in vivo*), the complex physical, biological, and social processes associated with language learning, use, and knowledge may be abstracted into psycholinguistic theories describing operations on mental representations. In turn, the behaviors of these mental representations may be interpretable in terms of coherent linguistic structure. In neural networks (*in silico*), the complex patterns of weights and activations may be abstracted into interpretable circuits that approximately compute linguistically meaningful features like whether a word is singular and a grammatical subject. In turn, those interpretable circuits may be interpretable in terms of a larger abstraction of a coherent linguistic structure. Thus linguistic structure is a real pattern at the highest level of abstraction, which allows us to understand language as it is implemented by a human or by a neural network.

4.2. What LM interpretability can tell us about human language

We do not mean to suggest that linguists and cognitive scientists should identify compressible patterns in behavior, call it a real pattern, and declare victory. Perhaps *the* grand challenge of all of cognitive science is to uncover the internal mechanisms in neural networks and brains that underlie and implement behavioral patterns (Smolensky, 1988). Finding these internal mechanisms is not a prerequisite for the reality of linguistic structure – the mechanisms may be irreducibly complex. But when mechanisms *are* found in neural networks, they can be informative not only about linguistic structure itself, but also about how language processing and learning may be implemented in the human brain.

4.2.1. Interpretability in an idealized thought experiment

Consider a thought experiment, using the linguistic concept of c-command (a particular relationship between elements in a linguistic tree which has been posited to be important for a variety of structure-sensitive linguistic patterns). In this thought experiment, we get embedded representations for each word and use an interpretability method to get a representation of the pairwise relationship between each pair of words in each sentence. We analyze these word pair representations, and we find that, in layer *n*, neuron *m* patterns in an interesting way. If and only if the value of that particular neuron is positive, then the first word in the pair c-commands the second word. If it's negative, then we know that it doesn't. Further experimentation reveals that this pattern is remarkably consistent. (N.B.: We want to be clear that no experiment would ever give such clean results; LMs just don't represent anything in such human-interpretable ways.)

If we did discover a result like this, we would be justified in drawing several conclusions. First, the ease with which we extracted the c-command relationship would be compelling evidence that, for performing linguistic tasks in English, understanding which words c-command which other words is useful – so useful that the model learned it emergently. Second, we would conclude that learning the c-command relationship is possible with relatively little built-in language-specific bias. We would thus update our credence that the learning of complex syntactic relationships like c-command is possible without a built-in UG. Third, we could use this method to test various competing theories about which kinds of structural relationships should be more or less prominently represented in models. Fourth and most importantly, we could use these results to try to work out how an abstract structural relationship like c-command can be represented in the weights of any system which, like the brain, is composed of cascades of activating neurons. Such explorations would be informative as proof-of-concept *in silico* models of how the brain might process language.

If we *didn't* find the c-command relationship in our model representations, that's a trickier scenario. We might simply not have used the right methods to find c-command, even though it's robustly represented. Or it might turn out that, while c-command is critical to human processing, LMs process language sans c-command. Or it might turn out that neither humans nor models need to represent something like c-command to demonstrate proficient linguistic behavior. Even in this negative case, though, the results could be fruitful for generating hypotheses and stimulating further inquiry, grounded in data and requiring precise formulations.

Thus, while the linguistic conclusions we can draw from LM interpretability experiments require care, it should be relatively uncontroversial that the idealized c-command experiment above would be linguistically interesting – even while recognizing a wide variety of opinions as to whether the current slate of results from LM experiments *are* linguistically interesting.

4.2.2. Interpretability in the real world

Our c-command example is not realistic: like all complex systems, language models don't learn and represent information in neat and

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

interpretable ways. Rather, it takes a lot of work to ask questions about how LMs represent parts of speech (Tenney et al., 2019) or grammatical dependencies (Hewitt & Manning, 2019). And the answers that come out are messy.

But we are optimistic. The extent to which LMs are "black-boxes" is now overstated because of major advances in the field of neural network interpretability. For example, as discussed in Section 2.2, researchers now have a grip on how neural systems use the geometry of word embeddings to represent syntactic relations among words (Hewitt & Manning, 2019; Chi et al., 2020; Eisape et al., 2022; Diego-Simón et al., 2024).

More recent techniques have shown, in a mechanistic way, exactly how artificial neural circuits can be used to perform higher-level computations (Lindsey et al., 2025), finding neural circuits and induction heads in transformers (Elhage et al., 2021) or using sparse autoencoders to find key features in model representations (Huben et al., 2023). These techniques often depend on *causal* manipulation – showing that, if particular parts of the neural network are perturbed or patched in particular ways, the output is affected in predictable ways (Geiger et al., 2023). For instance, Wang et al. (2023) shows how it is possible to trace a circuit that controls the completion of an object (e.g., "When Mary and John went to the store, John gave a drink to ___." where the intended answer is "Mary"). Many papers have now used these techniques to *causally* study what parts of a network are responsible for particular kinds of complex linguistic behavior like grammatical number agreement (Lasri et al., 2022; Finlayson et al., 2021; Mueller et al., 2022; Lakretz et al., 2019), verb conjugation (Hao & Linzen, 2023), animacy processing (Hanna et al., 2023), and various kinds of long-distance dependencies (Arora et al., 2024). These techniques demonstrate how complex linguistic behavior can actually be implemented mechanistically in a neural system. These mechanisms are strong hypotheses for how syntactic relationships are represented in the human brain as well.

It is our hope – and also our prediction – that interpretability techniques will make it possible to close the gap between linguistic theory and implementation (mechanistic interpretability is a big part of "the way forward" per Millière & Buckner, 2024). This broad line of work, exploring symbolic representations in systems that solve genuinely interesting linguistic tasks, has started to make good on the promise of Smolensky's (1988) prescient vision "in which traditional and connectionist theoretical constructs collaborate intimately to provide an understanding of cognition".

4.3. LMs are a proof of concept for gradient, usage-based theories of language

The real patterns account leaves open the possibility that a wide variety of linguistic theories can be fruitfully combined with LMs. But there is a separate question raised here, which is whether the LM era should push us toward any particular theory for explaining language. We believe that the success of LMs strengthens the case for existing usage-based language theories based on gradient representations. By their very nature, such theories have been difficult to formalize to the extent that one could, for example, generate or process sentences with them. LMs serve as model systems for how these ideas can be implemented, and as proof of concept that they can work.

4.3.1. Word meanings

Traditional views in semantics often posit that words have discrete meanings and can be combined using systematic rules (Heim & Kratzer, 1998). In contrast, a distributional view of meaning assumes meanings are best captured by usage patterns (Potts, 2019; Goldberg, 2019; Erk, 2012; Baroni & Lenci, 2010; Baroni et al., 2014; Elman, 1990a), an intuition now operationalized inside language models as continuous high-dimensional vectors that encode the contexts in which a word appears, resulting in a rich web of lexical relationships represented within the geometry of the word vector space (Mikolov et al., 2013). If we think of the meaning of the word *fire* as being represented by its embedding in context, we find that the meaning of the word is slightly different in every context in which it occurs, but clustering together in space (campfires in a distinct cluster from a building fire even though these meanings are in some theories often lumped in as the same "meaning"; Chronis & Erk, 2020). As such, LMs show that we might not need a strict separation between core meaning and a richly context-dependent and pragmatics-laden meaning in use. LMs give us a way of formalizing the intuition that meaning might be gradient, fuzzy, and usage-based in a way that works in practice, not just in theory.

4.3.2. Syntactic categories and grammatical roles

LMs seem to be able to perform tasks that require knowledge of syntactic category and grammatical role, such as generalizing part of speech when encountering a novel token (Kim & Smolensky, 2021). One might then be tempted to ask: what are the syntactic categories an LM uses? And we can, for instance, probe for grammatical category in models and see whether the categories that emerge correspond to ones posited by linguists (Chi et al., 2020; Papadimitriou et al., 2021)? A key finding of this work is that, while there are natural clusters in part of speech and grammatical role, such clusters are necessarily fuzzy and gradient. For instance, Papadimitriou et al. (2021) found robust representations of subjecthood. But they were graded, with intransitive subjects and passive subjects less "subject-y" than the subjects of transitive active verbs. These representations could be discretized to make judgments like "BERT represents this word as a subject but not this one," but this discretization would be lossy. This graded notion of syntactic category and grammatical role has a long history in functional linguistics (Ross, 1972) and, as such, can be useful in giving us a new way to think about claims that word classes are not stable across languages (Haspelmath, 2012) and that grammatical categories exist in graded hierarchies (Comrie, 1989; Keenan & Comrie, 1977).

4.3.3. Compositionality

While traditional theories have emphasized the compositional nature of language, the construction-based approach has emphasized that not all (or even most) language can be explained strictly compositionally (Goldberg, 2009). This position argues for a graded notion of compositionality: knowing what "green tea" is requires more than just knowing the meaning of "green" and the meaning of "tea" – even though this isn't strictly an idiom in that the meanings of "green" and "tea" are still relevant. There has been a lot of work on whether LMs can handle strict compositionality (Lake & Baroni, 2018; Kim & Linzen, 2020; Russin et al., 2024), as well as work on how LMs handle more idiosyncratic constructions (Mahowald, 2023; Misra & Mahowald, 2024; Weissweiler et al., 2023; Tseng et al., 2022; Devlin et al., 2019). Neural models can combine the best of both worlds in that they can maintain multiple levels of representation at once (Baroni et al., 2014). They don't have to decide whether "green tea" is fully composed or fully stored: it can be a little of both. They can even be used to measure

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

the degree of compositionality quantitatively (Rathi et al., 2021; Socolof et al., 2022).

4.3.4. Separation of linguistic layers

Traditional linguistic theory has often posited a cognitive separation in processing between linguistic layers (phonology, syntax, semantics, and pragmatics). But evidence from psycholinguistics (e.g., Tanenhaus et al., 1995; Shain et al., 2024; Fedorenko et al., 2020) suggests these layers are not processed in neat discrete stages but are richly integrated. LMs do seem to have some separation such that, for example, lower layers of the model handle morphology with higher layers handling syntax and semantics (Tenney et al., 2019). However, the nature of their architecture is to share information in hard-to-untangle nonlinear ways. So their functional separation is not strict but fuzzy and gradient – similar to psycholinguistic evidence for humans.

4.4. What the inductive biases of LMs can tell us about language

Since neural language models do have inductive biases, perhaps one reason for their success in learning language is that those inductive biases meaningfully align with linguistic structure. Indeed, a number of recent studies have investigated this possibility.

In one such study, Kallini et al. (2024) refute the claim that neural LMs can learn any language, including unnatural ones, equally well (Bolhuis et al., 2024). They compare the learning curves for the GPT-2 architecture trained on language modeling on English text against models trained on various transformations of the English text, designed to create languages that are intuitively "impossible," but which still have the same level of overall predictability as the original English text. For example, one transformation applies a deterministic shuffling function to the tokens of English text, creating extraordinarily complex but deterministic word order rules that violate all known formal characterizations of syntax, and another transformation creates a new agreement marker that must appear exactly 4 tokens away from a verb, also an unnatural pattern. Kallini et al. (2024) find that the model learns from real English text consistently faster than these baselines (see also Mitchell & Bowers, 2020; Yang et al., 2025; Xu et al., 2025; Ziv et al., 2025, among others).

These results and others show that transformers and related models have inductive biases that align with human language. However, the major determinant of inductive biases in LM is not that they are restricted to a particular formal language class, as might be expected from the generative linguistics paradigm. In fact, in terms of formal expressivity, it seems that transformers are mismatched with the usual formal language classes used to characterize language. Whereas human language is sometimes characterized using (extensions of) the Chomsky–Schützenberger hierarchy (Chomsky & Schützenberger, 1963; Vijay-Shanker et al., 1987; Weir, 1988), which encompasses well-known classes such as regular and context-free languages, modern transformers as they are currently applied (and other recently successful language model architectures such as State Space Models: Gu & Dao, 2024) seem to inhabit formal language classes defined by *circuit complexity* (Merrill et al., 2022; Strobl et al., 2024; Merrill et al., 2024), a formal language hierarchy which is orthogonal to the Chomsky–Schützenberger hierarchy.[13] To the extent that transformers have inductive biases that help (or hinder) their language learning, they likely arise from something other than the expressive limits of their architecture.

Below we consider two apparent learning biases of modern LMs which may be aligned with the structure of human language: information locality and low sensitivity.

4.4.1. Information locality

Human languages are structured in a way such that elements that statistically predict each other are usually close to each other. For example, in phrase such as *big brown box*, the noun *box* and the adjective *brown* are highly predictive of each other – boxes, especially cardboard ones, are often brown, for many reasons – and so these words are likely to be close to each other: the alternate order *brown big box* sounds odd or like it is conveying some other special meaning (Futrell, 2019; Culbertson et al., 2020; Scontras, 2023; Dyer et al., 2023). Locality ideas of this kind pervade human language (Behaghel, 1930; Givón, 1991; Futrell, 2019; Mansfield, 2021; Hahn et al., 2021a; Mansfield & Kemp, 2023): words are usually contiguous units, usually modified by prefixes and suffixes (directly adjacent to them, ordered by "relevance" to the root: Bybee, 1985; Saldana et al., 2024), and words linked by syntactic dependencies tend to be close to each other

(Gibson, 1991, 1998; Liu, 2008; Liu et al., 2017), more than would expected under random grammars within a linguistically realistic formalism (Gildea & Temperley, 2007; Park & Levy, 2009; Gildea & Temperley, 2010; Futrell et al., 2015, 2020b).

Autoregressive LMs such as GPT-2 also show a bias toward information locality, as demonstrated by several of the experiments of Kallini et al. (2024) (discussed above): many of the counterfactual languages which are harder to learn in those experiments are also those that disrupt information locality. The bias toward locality seems to come from the next-token prediction task performed by LMs. The bias toward information locality is an "ember of autoregression" in the terminology of McCoy et al. (2023), one which helps language learning and is likely shared with humans.

4.4.2. Relatively low sensitivity

Another related inductive bias in transformers is the bias toward learning functions with low sensitivity or low polynomial degree (Hahn et al., 2021b; Abbe et al., 2023; Bhattamishra et al., 2023). Sensitive functions are functions on input strings whose outputs change drastically based on small changes to the input. For example, a function on input bitstrings that counts the parity of the input – returning odd or even depending on the number of 1's in the input – is maximally sensitive and high-degree, because any single change to the input bits will flip the output of the function (O'Donnell, 2014). Although the Transformer architecture has the ability to represent highly sensitive functions in terms of its representational capacity, this turns out not to be relevant to its inductive bias in learning. Rather, the bias toward low-sensitivity functions comes from the shape of the loss landscape induced by the model. In particular, any parameter setting representing a highly sensitive function in the Transformer architecture must be *brittle*, meaning that a small change to the parameters would make the Transformer produce some different, lower-sensitivity function (Hahn & Rofin, 2024). Thus, high-sensitivity functions are unlikely to be reached through a gradient-descent-based learning process.

Human languages, viewed (for example) as functions from strings to meanings or to grammaticality judgments, also seem to be *relatively* low sensitivity (Hahn et al., 2021b). We do not find

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

human languages where, for example, a word is well formed if and only if it contains an odd number of some certain segment, even though it would be easy to express this language using a regular grammar.

However, like information locality, low sensitivity is not an absolute formal restriction on languages. For example, calculating the meaning of iterated negation is like a parity function: What does it mean when someone says they are *not not not not not not* wearing a tie? These high-sensitivity parts of language are nonetheless rare in usage and difficult to understand in practice. Relative low sensitivity is an important statistical property of human language, perhaps reflecting a general cognitive constraint for humans, which has come to light as the result of taking the inductive biases of LM architectures seriously.

4.5. Convergence of LMs and psycholinguistics on predictive processing

It is not a coincidence that the basic paradigm for training language models – incremental prediction of upcoming input – mirrors ideas about how human language processing works. In early neural network work, Elman (1990b) cites two reasons for choosing to use a prediction task in training his models: "minimal assumptions about special knowledge required for training" and the fact that "much of what listeners do involves anticipation of future input". Indeed, there is a deep functional convergence between the predictive approach to language modeling and the ways that human language is processed and structured.

Elman's statement about what listeners do is not speculative. Experimental psycholinguistics has delivered a picture of human language processing that is deeply entwined with the task of autoregressive prediction, mirroring other areas of cognition (Rao & Ballard, 1999; Bar, 2009; Friston & Kiebel, 2009; Ryskin & Nieuwland, 2023). In comprehension (the process of decoding meaning from linguistic form in real time), the way that human processing works is influenced by the predictability of each word (or sub-word unit) conditional on previous words (Hale, 2001; Levy, 2008; Wilcox et al., 2020, 2023b; Xu et al., 2023). Furthermore, language comprehension is highly *incremental* (Tanenhaus et al., 1995; Smith & Levy, 2013), meaning that information of different kinds is processed and integrated as quickly as possible as it comes in. There is also a high degree of incrementality in language production (Ferreira & Swets, 2002), and incremental predictability plays a key role there too: for example, disfluencies and pauses in speech happen before unpredictable words (Goldman-Eisler, 1957; Beattie & Butterworth, 1979; Dammalapati et al., 2019; Futrell, 2023), and unpredictable words are enunciated more slowly (Bell et al., 2009; Upadhye & Futrell, 2025).

Many of these studies about the effects of predictability in language processing would have been impossible without large language models. Even after converging on the idea that prediction is an important part of language processing, psycholinguistics as a field was hampered by the fact that strong probabilistic models for language were not available, resulting in conclusions based on what now seem to be fairly weak models and methods of estimating incremental probability (Frank & Bod, 2011; Fossum & Levy, 2012; Smith & Levy, 2013). Now, neural models make it possible to probe the limits of predictive models of language processing (Wilcox et al., 2020, 2023b; Xu et al., 2023), including the discovery that certain phenomena seem to go beyond the simplest predictive models (van Schijndel & Linzen, 2018, 2021; Huang et al., 2024; Staub, 2024), and that human processing is better characterized by prediction under resource constraints similar to but more severe than those present in neural models (Futrell et al., 2020a; Hahn et al., 2022; Kuribayashi et al., 2022; Timkey & Linzen, 2023; Oh & Schuler, 2023; De Varda & Marelli, 2024; Clark et al., 2025).

4.6. Functional explanations for human language

LMs also orient us toward another route to explanatory adequacy in linguistic theory, one which posits that the form of language is related to its function: communication of thought and social coordination under general cognitive constraints on how language is produced and comprehended (Chomsky, 2005; Gibson et al., 2019; Levshina, 2022; Bickel et al., 2024). The *functionalist* school of linguistics, which is often contrasted to the generative or formalist approach (Newmeyer, 1998), holds that the structure of language ultimately reflects constraints and pressures arising from the way language is actually used (Hawkins, 1994, 2004, 2014; Haspelmath, 2008; Comrie, 1989) – perhaps motivated by factors such as a language's "niche" such that languages with larger communities of speakers or more second language learners might have different pressures (Lupyan & Dale, 2010; Raviv et al., 2019).

We believe that linguistics and language modeling can make (and in fact already have made) fruitful contact at this functional level of explanation. By analogy, bird wings and airplane wings are very different things, but they share the *function* of flying in the Earth's atmosphere, and so they are both shaped by the constraints of aerodynamics (Marr, 1982; Gill, 1995). Similarly, neural language models and human language processing mechanisms are different things, but at the level of function, they both encode and decode information in linguistic strings *incrementally* and *predictively*. To the extent that human language is structured in a way that supports this kind of incremental prediction (e.g., through information locality), it is also aligned with the strengths and inductive biases of language models.

Indeed, language models have already proved a key tool in functionalist models of why languages are the way they are. A sizable literature exists about what kinds of languages emerge in simulated populations of agents who communicate by encoding and decoding meanings into strings using various neural architectures, and what constraints (on the environment, the agents, or the task) are necessary for the emergent languages to resemble natural language (Lazaridou et al., 2017; Mordatch & Abbeel, 2018; Steinert-Threlkeld, 2020; Kuciński et al., 2021; Chaabouni et al., 2021). For example, Hahn et al. (2020) show that certain universal properties of word order (Greenberg, 1963; Dryer, 1992) can be derived by finding grammars that optimize (1) the ease of recovering a parse tree from a string and (2) the ease of incrementally predicting each word, with both factors operationalized using neural networks (see also Kuribayashi et al., 2024). Clark et al. (2023) show that natural language word order seems to be structured in a way that minimizes the *variance* of word-by-word surprisals, again measured using neural LMs, in keeping with the theory of Uniform Information Density (Fenk & Fenk, 1980; Levy & Jaeger, 2007; Jaeger, 2010; Jaeger & Tily, 2011). More generally, because human language processing is highly probabilistic and predictive, theories of linguistic structure based on functional constraints must be evaluated using a strong probabilistic predictive model. Language models provide exactly that.

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

4.7. Upshots for linguistics beyond language structure

In much of this piece, we have focused on the upshot of LMs for linguistic structure, particularly morphosyntax. We did so for two main reasons: first, because of the centrality of these topics within linguistics over the last 60+ years; and, second, because mastery of linguistic *form* emerged in models earlier than other high-level abilities (Mahowald et al., 2024). Taken together, these developments have led to important insights about human language learning and processing.

But LMs have also made recent striking gains in domains like reasoning, logic, and long-form dialog. These abilities seem to emerge not just from pretraining, but from techniques like supervised fine-tuning, instruction-tuning, and reinforcement learning from human feedback whereby models are given specific feedback to make them more aligned with human behavior on specific tasks (Ouyang et al., 2022; Bai et al., 2022; Achiam et al., 2023). The importance of inference-time compute – whereby models learn in-context and perform additional computation during the process of generating text – has also become apparent for tasks like math and reasoning (OpenAI et al., 2024; Marjanović et al., 2025).

As such, just as it is fruitful to explore what inductive biases and representational mechanisms underlie the emergence of syntax in LMs, it is increasingly fruitful to study what inductive biases and structures are necessary and/or sufficient for other abilities in LMs like reasoning (Marjanović et al., 2025), theory of mind (Hu et al., 2025), planning (Liu et al., 2023), and other aspects of higher-level cognition. Many of our same arguments hold in these domains as well: the upshot of the modern view of learning, the Contravariance Principle, the role of the statistical tradition, real patterns.

Moreover, as LMs continue to develop, they increasingly are becoming part of our speech communities. Millions of humans regularly interact with chatbots, in some cases carrying on long extended dialogs with chatbots playing various roles. And, increasingly, text read by humans is written or co-written by AI. How will the presence of AI change language? How should we think about dialogs between human and non-human interlocutors? What social consequences will emerge from this new paradigm? Linguistics as a field, particularly sociocultural linguistics and related disciplines, is well positioned to be at the forefront of answering these questions (Bucholtz & Hall, 2005; Meyerhoff, 2006).

5. Conclusion

As Norm and Claudette would agree, this is a remarkable and pivotal moment in linguistics. Even ten years ago, it was not obvious that Claudette's dream of models that produce fluent and coherent text would be possible in her lifetime – or ever. It was also not obvious that, if it were to happen, it would happen using statistical systems trained largely on next-word prediction, using massive amounts of data. These systems don't use Norm's handcrafted rules. They don't rely on insights from generative linguistic theory, although they do draw on decades of linguistic work in distributional semantics and statistical language learning.

Rather than resisting this development, we contend it offers a golden opportunity for linguistics by enabling new kinds of research and opening up vistas of new hypotheses, methods, and research questions. Some of these questions will be new, like what kinds of artificial neural architectures best capture language, what kinds of biases they hold, and what the sources of those biases are. But we also have highlighted ways in which these methods can make progress on some of the oldest and most venerable questions in linguistics: what must be true of the input data for certain structures to be learned in principle, what are the constraints on what languages are possible, and how does the form of language relate to its function.

In response to Piantadosi's claim that language models refute Chomsky, Kodner et al. (2023) wrote a piece skeptical of the role of LMs in linguistic theory called "Why linguistics will thrive in the 21st century". We would go a step further and argue that, with the 21st century a quarter over, linguistics *is* thriving. But we think it's doing so not in spite of language models, but in part because of them and the opportunities they open up for new research directions and the light they shed on old questions.

Therefore, we reject the dichotomy in the discourse around Piantadosi's (2023) paper, which often seems to pit language models against linguistics. Linguistic structure, as described in linguistic theories, is real and important even if language models learn those structures emergently in a complex statistical way. Conversely, language models can point the way to new ways of thinking about the learning, processing, and fundamental nature of language.

But we do think there are some major revisions to some earlier accepted dogma that are warranted, including:

- Much of the field of linguistics has assumed that the form of language must be explained in terms of a symbolic formalism that constrains the forms of possible languages, and that this grammar formalism must represent innate human constraints on language, and that these constraints are logically necessary for language to be learned. The success of machine learning models should orient us away from this paradigm, because it shows that this approach is not necessary to characterize learning, and it opens up a new universe of statistical, quantitative, and functional theories to constrain the forms of possible languages and explain why only humans have language, while at the same time providing the tools to test those theories.
- Relatedly, the idea that the formal structure of linguistic competence should be the main focus of theoretical linguistics, which precedes other forms of study, is no longer tenable – language models reveal and allow us to characterize new dimensions of language such as how corpora of text express world knowledge, how the structure of usage supports learning or doesn't, and how fundamental information-processing constraints shape the way that language is represented in brains and machines.
- The success of machine learning models shows us how language can be represented in ways that are graded, probabilistic, and fuzzy, which should move us away from an insistence on discrete categorical frameworks. They also reveal potential soft constraints on the structure of language.
- The ontological basis for linguistic structure need not lie in an innate genetic endowment. Linguistic structure is a real pattern, just as real and worthy of study whether it lies innate in the human genome or whether it is learned entirely through inductive statistical learning with domain-general biases. Linguistic phenomena do not become less interesting when they are learnable in this way, they become *more* interesting.

More broadly, language models show that linguistics needs the rest of science. We will not answer the most fundamental questions

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

about language by focusing *only* on language, using *only* methods that are tailor-made for linguistics. Just as neural networks revealed principles that are now fundamental in our understanding of vision, they are now in a position to revolutionize our understanding of language. Just as physics was revolutionized by thermodynamic ideas which arose from the ad-hoc analysis of complex machines, linguistics stands to benefit from new ways of thinking about computation arising from the analysis of neural networks. There is now massive intellectual activity spurred by language models in cognitive science, philosophy, physics, statistical learning theory, and information theory. The science of language can draw ideas, methods, and inspiration from all of this by maintaining a spirit of deep, curious, open-minded engagement and integration. Human language is in many ways unique as a natural phenomenon, and this means that the science of language should integrate with other fields of science *more*, in order to find deeper unities and to delineate and explain how and why language is special.

In particular, we are optimistic about significantly revising what Stabler (1983) said in doubting that "any close connections between linguistic theory and biology will be forthcoming". He continues: "it is a crucial advantage of the computational approach that it has a functionalist vocabulary that does not require a type reduction to physicalistic concepts [. . .] since this seems to be required in linguistics and in psychological accounts of language processing and visual perception". We agree that there is a crucial advantage to the functionalist vocabulary used in linguistic theory. But by analogy to recent work in vision, we are optimistic about the role that artificial neural models will play in bridging the gap between theory and biology.

This progress will come through an expansive linguistics – expansive in the breadth and diversity of languages it considers, expansive in its methods, and expansive in its connections to related fields. Baroni (2022) lamented that, as of a 2021 exploration of citation records, Linzen et al.'s seminal work on subject–verb agreement was largely uncited within the linguistics community. But, since then, there are increasingly researchers using language models to ask questions that are informed by and which can inform linguistic theory.

On the computational side, while there is a persistent stereotype that the NLP community cares mostly about engineering, we have found there to be significant interest in fundamentally linguistic questions. Of the 7 papers that won Best Paper Awards at ACL 2024 (the largest annual conference focused on NLP), 1 was a direct response to a claim by Chomsky about language learning (Kallini et al., 2024), 1 was about inductive biases in models of relevance to questions about constraints on language (Hahn & Rofin, 2024), 1 was a theoretically motivated paper studying satisfiability in natural language of relevance to old questions about the complexity of language (Madusanka et al., 2024), 1 was a new method for measuring memorization in models and is of relevance for studying trade-offs in memorization vs. generalization in natural language (Lesci et al., 2024), 2 were about reconstructing or recovering ancient languages (Lu et al., 2024; Guan et al., 2024), and 1 presented an open-access multilingual model for broadening coverage of under-resourced languages (Üstün et al., 2024). Every one of these papers has relevance to scientific questions about human language, and in most cases one or more of the authors works in a linguistics department and/or has a degree in linguistics.

Thus, we don't see ourselves as calling for radical change, but rather as acknowledging the interdisciplinary work that is already thriving in the 21st century. Linguistically informed computational work is increasingly taking place within linguistics departments, where computational researchers are working alongside syntacticians, semanticists, phonologists, language documentation experts, sociocultural linguists, and experts in a wide variety of languages and language families. We think this is a very good development – one that both Norm and Claudette should embrace.

Acknowledgements. We thank Chris Potts for pointing out that he independently used a similar title in a talk called "Linguists for Deep Learning; or: How I Learned to Stop Worrying and Love Neural Networks" and encouraging us to use our title anyway. We thank David Beaver, Joan Bresnan, Ted Gibson, Greg Hickok, Laura Kalin, Andy Kehler, Robbie Kubala, Harvey Lederman, Connor Mayer, Kanishka Misra, Andrea Moro, Lisa Pearl, Jon Rawski, Greg Scontras, Alex Warstadt, Steve Wechsler, Andrew Wilson, Xin Xie, and three anonymous reviewers for helpful comments.

Financial support. KM was supported by NSF CRII grant 2104995 and NSF CAREER grant 2339729 from the Directorate of STEM Education (EDU) Division of Research on Learning in Formal and Informal Settings (DRL). RF was supported by NSF CRII grant 1947307.

Competing interests. KM has done contract work for OpenAI, unrelated to the content of this manuscript.

Notes

1 Nor is it necessarily sensible to evaluate models based on simply prompting them and asking whether sentences are grammatical or not (Hu & Levy, 2023) (as done by Dentella et al., 2023). To see this, consider that many linguists have held that grammatical knowledge can be represented by a formal system like a context-free grammar; even if a context-free grammar is a good model of linguistic knowledge, it could not be prompted in natural language to respond about whether a sentence is grammatical or not. Grammatical knowledge does not imply that a system can make explicit grammaticality judgments when prompted in natural language.

2 It has been objected that, while humans make errors in verb agreement and other syntactic constructions, they have an underlying error-free competence which may be distinguished from their errorful performance (Chomsky, 1965, Ch. 1), whereas the neural networks only have errorful performance with no underlying competence (Fox & Katzir, 2024). We believe this argument is mistaken for two reasons. First, it is indeed possible to define (multiple) competence–performance distinction(s) in neural networks, for example, by separating the knowledge contained in embeddings versus the performance of a linear decoder in extracting that knowledge (Pimentel et al., 2020; White et al., 2021; Csordás et al., 2024). Second, even for humans, we only ever have access to performance data (including subjective grammaticality judgments), and we infer mental constructs such as competence, I-language, etc., on the basis of that data (Piantadosi & Gallistel, 2024). We may make the same inferences based on neural network performance data. Competence is not a datum to be explained, but rather a theoretical construct that is useful for explaining performance data —it is a real pattern (Dennett, 1989; Nefdt, 2023) in the sense we will discuss in Section 4.1.

3 We note that these kinds of criticisms are harder to apply to studies based on more carefully controlled sets of sentences (for example, Marvin & Linzen, 2018; Futrell et al., 2019b; Hu et al., 2020; Wilcox et al., 2023a).

4 There remains controversy over the extent to which deep neural networks are the best model of human vision (see Bowers et al., 2023, for a more skeptical take along with spirited replies), but even skeptics of neural models judge that this research program has been fruitful.

5 We use the term inductive bias here to refer to anything that causes a learner to prefer one hypothesis over another, beyond any function of input data (Mitchell, 1980). Such biases exist and are necessary even in learning algorithms that are not strictly 'inductive', and they may or may not be identifiable as Bayesian priors.

6 For example, under this view, it would be justified to criticize a theory of grammar such as Head-Driven Phrase Structure Grammar (HPSG: Pollard &

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

Sag, 1994; Sag et al., 2003) on the basis that it is Turing-complete, capable of generating any recursively enumerable formal language.

7 Given the centrality of this reasoning to much of linguistic theory, one might expect it to be backed up with strong experimental evidence that humans cannot or do not learn languages which violate the putatively universal principles of human languages. This is not the case: there is only limited and ambiguous experimental evidence for hard formal limits on human language learning.

Among the most on-target existing studies is Smith et al. (1993), whose object of study was a man described as a polyglot savant living in a mental health facility. This individual and four control subjects (linguistics undergraduates) were tasked with learning artificial languages designed to be 'impossible' in three ways: (1) negation and tense are indicated by word order, (2) there is an agreement pattern judged to be impossible, and (3) the position of an emphatic marker is determined by a rule involving counting words. Results are not systematically reported, but seem to indicate that the polyglot was able to learn the 'impossible' word order and agreement rules (1) and (2), but not the rule for the emphatic marker (3).

Another commonly cited experimental work on this topic is Musso et al. (2003), who expose German speakers to Italian and Japanese sentences, either following the real rules of those languages, or following modified rules deemed to be linguistically impossible, for example placing a negation marker after the third morpheme from the beginning of a sentence. The result is equally accurate learning of the 'natural' and 'unnatural' languages. fMRI on the subjects shows that the real languages elicit activity in the left inferior frontal gyrus, while the unnatural ones elicit activity elsewhere. We believe the meaning of these results is unclear. Only a small number of languages and participants (all of whom were already native speakers of largely hierarchically-structured languages) were tested, the localization of syntax in the brain is still contentious, and the patterns of brain activity for the 'unnatural' languages might reflect a lack of practice with such patterns, rather than their impossibility or a qualitative difference between linear and hierarchical rules.

8 See, for example, Kodner et al.'s (2022) criticism of Yang and Piantadosi's (2022) model of language learning as Bayesian program induction, a model which successfully learns grammars of various formal classes given small amounts of string input, thus addressing the Poverty of the Stimulus problem for formal structures. Although the model readily meets the challenge of inducing formal structure from strings, it has been dismissed by some in the linguistics literature because the same model could also learn grammars that are unlike human language.

9 Although these are common examples, it is not clear exactly how generative formalisms rule out the unnatural hypotheses here. In particular, the languages implied by the unnatural hypotheses are context-free, just as much as the languages implied by the natural hypotheses. So these (string) languages could be generated from, for example, Minimalist Grammars (Chomsky, 1993; Stabler, 1997), since Minimalist Grammars generate a superset of context-free languages (Michaelis, 1998).

10 In certain settings and with the right regularization, the best performance comes from models whose capacity is poised right at the point where training data and model parameters are balanced—the right regularization in these settings avoids the sharp spike in loss characteristic of double descent (Maloney et al., 2022, §4.2).

11 Indeed, the analysis by Mollica and Piantadosi (2019) suggests that syntactic structure makes up only a very small portion of the information necessary to learn a language.

12 Early generative linguistics went further in its goals, claiming that the structures identified through generative analysis would not only provide a characterization of language, but also be a necessary and explicit part of any theory of how language could be learned and processed (Chomsky, 1965, 1971; Bresnan, 1982). See Stabler (1983) for discussion of what he sees as confusion within the field as to whether or not generative linguistic theories are intended to be theories of the representations used by the brain during processing, as opposed to theories that constrain possible language. He concludes that Chomsky and others often conflated theories of grammar and theories of mental representation and processing, but that this position was not well grounded and that "it seems unlikely that linguists and psychologists really want to claim any such thing".

13 Interestingly, circuit complexity has also been used to characterize the computational capacity of biologically realistic populations of neurons (Maass, 1997; Maass & Markram, 2004). So human performance may also be ultimately limited in this way.

References

Abbe, E., Bengio, S., Lotfi, A., & Rizk, K. (2023). Generalization on the unseen, logic reasoning and degree curriculum. *Proceedings of the International Conference on Machine Learning*, *202*, 31–60.

Abney, S. (1996). Statistical methods and linguistics. In *The balancing act: Combining symbolic and statistical approaches to language* (pp. 1–26). MIT Press.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Adam, S. P., Alexandropoulos, S.-A. N., Pardalos, P. M., & Vrahatis, M. N. (2019). No free lunch theorem: A review. In *Approximation and optimization: Algorithms, complexity and applications* (pp. 57–82). Springer.

Adger, D. (2003). *Core syntax: A minimalist approach*. Oxford University Press.

Adger, D. (2022). What are linguistic representations? *Mind & Language*, *37*(2), 248–260.

Ahuja, K., Balachandran, V., Panwar, M., He, T., Smith, N. A., Goyal, N., & Tsvetkov, Y. (2025). Learning syntax without planting trees: Understanding hierarchical generalization in transformers. *Transactions of the Association for Computational Linguistics*, *13*, 121–141.

Arora, A., Jurafsky, D., & Potts, C. (2024). CausalGym: Benchmarking causal interpretability methods on linguistic tasks. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd annual meeting of the Association for Computational Linguistics (volume 1: long papers)* (pp. 14638–14663). Association for Computational Linguistics.

Arora, S., Liang, Y., & Ma, T. (2017). *A simple but tough-to-beat baseline for sentence embeddings*. *International conference on learning representations*, Toulon, France.

Attias, I., Dziugaite, G. K., Haghifam, M., Livni, R., & Roy, D. M. (2024). Information complexity of stochastic convex optimization: Applications to generalization, memorization, and tracing. *Proceedings of the International Conference on Machine Learning*, *81*, 2035–2068.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*.

Bar, M. (2009). Predictions: A universal principle in the operation of the human brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1181–1182.

Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. In W. A. Rosenblith (Ed.), *Sensory communication* (pp. 217–233). MIT Press.

Barlow, H. B. (1989). Unsupervised learning. *Neural Computation*, *1*(3), 295–311.

Baroni, M. (2022). On the proper role of linguistically oriented deep net analysis in linguistic theorising. In *Algebraic structures in natural language* (pp. 1–16). CRC Press.

Baroni, M., Bernardi, R., & Zamparelli, R. (2014). Frege in space: A program for compositional distributional semantics. *Linguistic Issues in Language Technology*, *9*, 241–346.

Baroni, M., & Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, *36*(4), 673–721.

Baxter, J. (2000). A model of inductive bias learning. *Journal of Artificial Intelligence Research*, *12*, 149–198.

Beattie, G. W., & Butterworth, B. L. (1979). Contextual probability and word frequency as determinants of pauses and errors in spontaneous speech. *Language and Speech*, *22*(3), 201–211.

Behaghel, O. (1930). Zur Wortstellung des Deutschen. *Language*, *6*(4), 29–33.

Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, *116*(32), 15849–15854.

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, *60*(1), 92–111.

Belth, C. A., Payne, S. R., Beser, D., Kodner, J., & Yang, C. (2021). The greedy and recursive search for morphological productivity. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*, 2869–2875.

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics* (pp. 5185–5198). Association for Computational Linguistics.

Bernardy, J.-P., & Lappin, S. (2017). Using deep neural networks to learn syntactic agreement. *Linguistic Issues in Language Technology*, *15*(2), 1–15.

Berwick, R. C., Pietroski, P., Yankama, B., & Chomsky, N. (2011). Poverty of the stimulus revisited. *Cognitive Science*, *35*(7), 1207–1242.

Bhattamishra, S., Patel, A., Kanade, V., & Blunsom, P. (2023). Simplicity bias in Transformers and their ability to learn sparse Boolean functions. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the Association for Computational Linguistics (volume 1: long papers)* (pp. 5767–5791). Association for Computational Linguistics.

Bialek, W., Van Steveninck, R. R. D. R., & Tishby, N. (2006). Efficient representation as a design principle for neural coding and computation. In *2006 IEEE international symposium on information theory* (pp. 659–663). IEEE.

Bickel, B., Giraud, A.-L., Zuberbühler, K., & van Schaik, C. P. (2024). Language follows a distinct mode of extra-genomic evolution. *Physics of Life Reviews*, *50*, 211–225.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Blasi, D., Anastasopoulos, A., & Neubig, G. (2022). Systematic inequalities in language technology performance across the world's languages. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th annual meeting of the Association for Computational Linguistics (volume 1: long papers)* (pp. 5486–5505). Association for Computational Linguistics.

Bloomfield, L. (1926). A set of postulates for the science of language. *Language*, *2*(3), 153–164.

Bolhuis, J. J., Crain, S., Fong, S., & Moro, A. (2024). Three reasons why AI doesn't model human language. *Nature*, *627*(8004), 489–489.

Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., Puebla, G., Adolfi, F., Hummel, J. E., Heaton, R. F., et al. (2023). Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, *46*, e385.

Bresnan, J. (1982). Introduction: Grammars as mental representations of language. In *The mental representation of grammatical relations* (pp. xvii–lii). MIT Press.

Bresnan, J., Cueni, A., Nikitina, T., & Baayen, H. (2007). Predicting the dative alternation. In *Cognitive foundations of interpretation* (pp. 69–94). Royal Netherlands Academy of Science, Amsterdam.

Bresnan, J., Dingare, S., & Manning, C. D. (2001). Soft constraints mirror hard constraints: Voice and person in English and Lummi. In *Proceedings of the LFG 01 conference* (pp. 13–32). CSLI Publications.

Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Lafferty, J., Mercer, R. L., & Roossin, P. S. (1990). A statistical approach to machine translation. *Computational Linguistics*, *16*(2), 79–85.

Bucholtz, M., & Hall, K. (2005). Identity and interaction: A sociocultural linguistic approach. *Discourse studies*, *7*(4–5), 585–614.

Bybee, J. L. (1985). *Morphology: A study of the relation between meaning and form*. John Benjamins, Amsterdam.

Bybee, J. L., & Hopper, P. (2001). *Frequency and the emergence of linguistic structure, volume 45*. John Benjamins Publishing Company.

Cao, R., & Yamins, D. (2021). Explanatory models in neuroscience: Part 1–taking mechanistic abstraction seriously. *arXiv preprint arXiv:2104.01490*.

Carnot, S. (1824). *Réflexions sur la puissance motrice du feu et sur les machines propres à développer cette puissance*. Bachelier, Paris.

Caucheteux, C., Gramfort, A., & King, J.-R. (2023). Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature Human Behaviour*, *7*(3), 430–441.

Chaabouni, R., Kharitonov, E., Dupoux, E., & Baroni, M. (2021). Communicating artificial neural networks develop efficient color-naming systems. *Proceedings of the National Academy of Sciences*, *118*(12), e2016569118.

Chang, T. A., & Bergen, B. K. (2022). Word acquisition in neural language models. *Transactions of the Association for Computational Linguistics*, *10*, 1–16.

Chater, N., Tenenbaum, J., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, *10*(7), 287–291.

Chater, N., & Vitányi, P. (2007). Ideal learning of natural language: Positive results about learning from positive evidence. *Journal of Mathematical Psychology*, *51*(3), 135–163.

Chesi, C. (2025). Is it the end of (generative) linguistics as we know it? *Italian Journal of Linguistics*, *37*(1), 3–44.

Chi, E. A., Hewitt, J., & Manning, C. D. (2020). Finding universal grammatical relations in multilingual BERT. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the Association for Computational Linguistics* (pp. 5564–5577). Association for Computational Linguistics.

Chomsky, N. (1957). *Syntactic structures*. Walter de Gruyter.

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press.

Chomsky, N. (1971). *Problems of knowledge and freedom*. Fontana.

Chomsky, N. (1988). *Language and problems of knowledge: The Managua lectures*. MIT Press.

Chomsky, N. (1993). A minimalist program for linguistic theory. In K. Hale, & S. J. Keyser (Eds.), *The view from Building 20* (pp. 1–52). MIT Press.

Chomsky, N. (1995). *The minimalist program, volume 28*. Cambridge University Press.

Chomsky, N. (2005). Three factors in language design. *Linguistic Inquiry*, *36*(1), 1–61.

Chomsky, N. (2023). *Conversations with Tyler: Noam Chomsky*. Conversations with Tyler Podcast.

Chomsky, N., Roberts, I., & Watumull, J. (2023). Noam Chomsky: The false promise of ChatGPT. *The New York Times*. https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html

Chomsky, N., & Schützenberger, M. P. (1963). The algebraic theory of context-free languages. In P. Braffort & D. Hirschberg (Eds.), *Computer programming and formal languages* (pp. 118–161). North Holland.

Christiansen, M. H., & Chater, N. (2016). The Now-or-Never Bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*, e62.

Chronis, G., & Erk, K. (2020). When is a bishop not like a rook? When it's like a rabbi! Multiprototype BERT embeddings for estimating semantic relationships. In R. Fernández & T. Linzen (Eds.), *Proceedings of the 24th conference on computational natural language learning* (pp. 227–244). Association for Computational Linguistics.

Clark, C., Oh, B.-D., & Schuler, W. (2025). Linear recency bias during training improves transformers' fit to reading times. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, & S. Schockaert (Eds.), *Proceedings of the 31st international conference on computational linguistics* (pp. 7735–7747). Association for Computational Linguistics.

Clark, T. H., Meister, C., Pimentel, T., Hahn, M., Cotterell, R., Futrell, R., & Levy, R. (2023). A cross-linguistic pressure for Uniform Information Density in word order. *Transactions of the Association for Computational Linguistics*, *11*, 1048–1065.

Clausius, R. (1865). Über verschiedene, für die Anwendung bequeme Formen der Hauptgleichungen der mechanischen Wärmetheorie. *Annalen der Physik und Chemie*, *125*, 353–400.

Collins, J. (2024). The simple reason LLMs are not scientific models (and what the alternative is for linguistics). Preprint available at: https://ling.auf.net/lingbuzz/008026

Comrie, B. (1989). *Language universals and linguistic typology* (2nd ed.). University of Chicago Press.

Constantinescu, I., Pimentel, T., Cotterell, R., & Warstadt, A. (2025). Investigating critical period effects in language acquisition through neural language models. *Transactions of the Association for Computational Linguistics*, *13*, 96–120.

Croft, W. (2001). *Radical construction grammar: Syntactic theory in typological perspective*. Oxford University Press.

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

Csordás, R., Potts, C., Manning, C. D., & Geiger, A. (2024). Recurrent neural networks learn to store and generate sequences using non-linear representations. *arXiv preprint arXiv:2408.10920*.

Culbertson, J., Schouwstra, M., & Kirby, S. (2020). From the world to word order: Deriving biases in noun phrase order from statistical properties of the world. *Language*, *96*(3), 696–717.

Dammalapati, S., Rajkumar, R., & Agarwal, S. (2019). Expectation and locality effects in the prediction of disfluent fillers and repairs in English speech. In *Proceedings of the 2019 conference of the North American chapter of the Association for Computational Linguistics: Student research workshop* (pp. 103–109). Association for Computational Linguistics.

de Santillana, G. (1955). *The crime of Galileo*. University of Chicago Press.

De Varda, A., & Marelli, M. (2024). Locally biased transformers better align with human reading times. In T. Kuribayashi, G. Rambelli, E. Takmaz, P. Wicke, & Y. Oseki (Eds.), *Proceedings of the workshop on cognitive modeling and computational linguistics* (pp. 30–36). Association for Computational Linguistics.

Dennett, D. C. (1989). *The intentional stance*. MIT press.

Dennett, D. C. (1990). The interpretation of texts, people, and other artifacts. *Philosophy and Phenomenological Research*, *50*(S), 177–194.

Dentella, V., Günther, F., & Leivada, E. (2023). Systematic testing of three language models reveals low language accuracy, absence of response stability, and a yes-response bias. *Proceedings of the National Academy of Sciences*, *120*(51), e2309583120.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Association for Computational Linguistics.

Diego-Simón, P., D' Ascoli, S., Chemla, E., Lakretz, Y., & King, J.-R. (2024). A polar coordinate system represents syntax in large language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, & C. Zhang (Eds.), *Advances in neural information processing systems, volume 37* (pp. 105375–105396). Curran Associates, Inc.

Dryer, M. S. (1992). The Greenbergian word order correlations. *Language*, *68*(1), 81–138.

Dyer, C., Kuncoro, A., Ballesteros, M., & Smith, N. A. (2016). Recurrent neural network grammars. In *Proceedings of the 2016 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies* (pp. 199–209). Association for Computational Linguistics.

Dyer, W., Torres, C., Scontras, G., & Futrell, R. (2023). Evaluating a century of progress on the cognitive science of adjective ordering. *Transactions of the Association for Computational Linguistics*, *11*, 1185–1200.

Eisape, T., Gangireddy, V., Levy, R., & Kim, Y. (2022). Probing for incremental parse states in autoregressive language models. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022* (pp. 2801–2813). Association for Computational Linguistics.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., & Olah, C. (2021). A mathematical framework for transformer circuits. *Transformer Circuits Thread*. https://transformer-circuits.pub/2021/framework/index.html.

Elman, J. L. (1990a). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

Elman, J. L. (1990b). *Representation and structure in connectionist models*. The MIT Press.

Erk, K. (2012). Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, *6*(10), 635–653.

Evanson, L., Lakretz, Y., & King, J. R. (2023). Language acquisition: Do children and language models follow similar learning stages? In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 12205–12218). Association for Computational Linguistics.

Everaert, M. B., Huybregts, M. A., Chomsky, N., Berwick, R. C., & Bolhuis, J. J. (2015). Structures, not strings: Linguistics as part of the cognitive sciences. *Trends in Cognitive Sciences*, *19*(12), 729–743.

Fedorenko, E., Blank, I. A., Siegelman, M., & Mineroff, Z. (2020). Lack of selectivity for syntax relative to word meanings throughout the language network. *Cognition*, *203*, 104348.

Feng, L., Tung, F., Ahmed, M. O., Bengio, Y., & Hajimirsadegh, H. (2024a). Were RNNs all we needed? *arXiv preprint arXiv:2410.01201*.

Feng, S. Y., Goodman, N., & Frank, M. (2024b). Is child-directed speech effective training data for language models? In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 22055–22071). Association for Computational Linguistics.

Fenk, A., & Fenk, G. (1980). Konstanz im Kurzzeitgedächtnis—Konstanz im sprachlichen Informationsfluß. *Zeitschrift für experimentelle und angewandte Psychologie*, *27*(3), 400–414.

Ferreira, F., & Swets, B. (2002). How incremental is language production? Evidence from the production of utterances requiring the computation of arithmetic sums. *Journal of Memory and Language*, *46*(1), 57–84.

Finlayson, M., Mueller, A., Gehrmann, S., Shieber, S., Linzen, T., & Belinkov, Y. (2021). Causal analysis of syntactic agreement mechanisms in neural language models. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th annual meeting of the Association for Computational Linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)* (pp. 1828–1843). Association for Computational Linguistics.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930–1955. In J. R. Firth (Ed.), *Studies in linguistic analysis* (pp. 1–32). Basil Blackwell.

Flynn, M. (2013). The great Ptolemaic smackdown and down 'n dirty mud wrassle. *Analog*.

Fossum, V., & Levy, R. (2012). Sequential vs. hierarchical syntactic models of human incremental sentence processing. In *Proceedings of the 3rd workshop on cognitive modeling and computational linguistics* (pp. 61–69). Association for Computational Linguistics.

Fox, D., & Katzir, R. (2024). Large language models and theoretical linguistics. *Theoretical Linguistics*, *50*(1–2), 71–76.

Frank, S. L., & Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, *22*(6), 829–834.

Frazier, L., & Fodor, J. D. (1978). The sausage machine: A new two-stage parsing model. *Cognition*, *6*(4), 291–325.

Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*, 1211–1221.

Futrell, R. (2019). Information-theoretic locality properties of natural language. In *Proceedings of the first workshop on quantitative syntax (Quasy, SyntaxFest 2019)* (pp. 2–15). Association for Computational Linguistics.

Futrell, R. (2023). Information-theoretic principles in incremental language production. *Proceedings of the National Academy of Sciences*, *120*(39), e2220593120.

Futrell, R., Gibson, E., & Levy, R. P. (2020a). Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, *44*, e12814.

Futrell, R., Levy, R. P., & Gibson, E. (2020b). Dependency locality as an explanatory principle for word order. *Language*, *96*(2), 371–413.

Futrell, R., Mahowald, K., & Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, *112*(33), 10336–10341.

Futrell, R., Qian, P., Gibson, E., Fedorenko, E., & Blank, I. (2019a). Syntactic dependencies correspond to word pairs with high mutual information. In *Proceedings of the fifth international conference on dependency linguistics (Depling, SyntaxFest 2019)* (pp. 3–13). Association for Computational Linguistics.

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

Futrell, R., Wilcox, E., Morita, T., Qian, P., Ballesteros, M., & Levy, R. (2019b). Neural language models as psycholinguistic subjects: Representations of syntactic state. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies, volume 1 (long and short papers)* (pp. 32–42). Association for Computational Linguistics.

Gauthier, J., Hu, J., Wilcox, E., Qian, P., & Levy, R. (2020). SyntaxGym: An online platform for targeted evaluation of language models. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 70–76). Association for Computational Linguistics.

Geiger, A., Ibeling, D., Zur, A., Chaudhary, M., Chauhan, S., Huang, J., Arora, A., Wu, Z., Goodman, N., Potts, C., et al. (2023). Causal abstraction: A theoretical foundation for mechanistic interpretability. *arXiv preprint arXiv:2301.04709*.

Gibson, E. (1991). *A computational theory of human linguistic processing: Memory limitations and processing breakdown*. Carnegie Mellon University.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, *68*(1), 1–76.

Gibson, E., Futrell, R., Piantadosi, S. T., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. P. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences*, *23*(5), 389–407.

Gildea, D., & Temperley, D. (2007). Optimizing grammars for minimum dependency length. In *Proceedings of the 45th annual meeting of the Association for Computational Linguistics*, pages 184–191, Prague, Czech Republic.

Gildea, D., & Temperley, D. (2010). Do grammars minimize dependency length? *Cognitive Science*, *34*(2), 286–310.

Gill, F. B. (1995). *Ornithology*. Macmillan.

Givón, T. (1991). Isomorphism in the grammatical code: Cognitive and biological considerations. *Studies in Language*, *15*(1), 85–114.

Goldberg, A. E. (2009). *Constructions work*. Walter de Gruyter GmbH & Co. KG.

Goldberg, A. E. (2019). *Explain me this: Creativity, competition, and the partial productivity of constructions*. Princeton University Press.

Goldblum, M., Finzi, M. A., Rowan, K. & Wilson, A. G. (2024, July). Position: the no free lunch theorem, Kolmogorov complexity, and the role of inductive biases in machine learning. In *Forty-first International Conference on Machine Learning*.

Goldman-Eisler, F. (1957). Speech production and language statistics. *Nature*, *180*(4600), 1497–1497.

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., et al. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, *25*(3), 369–380.

Goyal, A., & Bengio, Y. (2022). Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A*, *478*(2266), 20210068.

Greenberg, J. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In *Universals of language* (pp. 40–70). MIT Press.

Gu, A., & Dao, T. (2024). Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*.

Guan, H., Yang, H., Wang, X., Han, S., Liu, Y., Jin, L., Bai, X., & Liu, Y. (2024). Deciphering oracle bone language with diffusion models. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd annual meeting of the Association for Computational Linguistics (volume 1: long papers)* (pp. 15554–15567). Association for Computational Linguistics.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. In M. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of the 2018 conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies, volume 1 (long papers)* (pp. 1195–1205). Association for Computational Linguistics.

Hahn, M., Degen, J., & Futrell, R. (2021a). Modeling word and morpheme order in natural language as an efficient tradeoff of memory and surprisal. *Psychological Review*, *128*(4), 726–756.

Hahn, M., Futrell, R., Levy, R., & Gibson, E. (2022). A resource-rational model of human processing of recursive linguistic structure. *Proceedings of the National Academy of Sciences*, *119*(43), e2122602119.

Hahn, M., Jurafsky, D., & Futrell, R. (2020). Universals of word order reflect optimization of grammars for efficient communication. *Proceedings of the National Academy of Sciences*, *117*(5), 2347–2353.

Hahn, M., Jurafsky, D., & Futrell, R. (2021b). Sensitivity as a complexity measure for sequence classification tasks. *Transactions of the Association for Computational Linguistics*, *9*, 891–908.

Hahn, M., & Rofin, M. (2024). Why are sensitive functions hard for Transformers? In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd annual meeting of the Association for Computational Linguistics (volume 1: long papers)* (pp. 14973–15008). Association for Computational Linguistics.

Hale, J. T. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American chapter of the Association for Computational Linguistics and Language Technologies* (pp. 1–8). Association for Computational Linguistics.

Hanna, M., Belinkov, Y., & Pezzelle, S. (2023). When language models fall in love: Animacy processing in transformer language models. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 12120–12135). Association for Computational Linguistics.

Hao, S., & Linzen, T. (2023). Verb conjugation in transformers is determined by linear encodings of subject number. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 4531–4539). Association for Computational Linguistics.

Harris, Z. S. (1951). *Methods in structural linguistics*. University of Chicago Press.

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(23), 146–162.

Harris, Z. S. (1955). From phoneme to morpheme. *Language*, *31*, 190–222.

Haspelmath, M. (2008). Parametric versus functional explanations of syntactic universals. In T. Biberauer (Ed.), *The limits of syntactic variation* (pp. 75–107). John Benjamins.

Haspelmath, M. (2009). Framework-free grammatical theory. In *The Oxford handbook of linguistic analysis*. Oxford University Press.

Haspelmath, M. (2012). How to compare major word-classes across the world's languages. *Theories of everything: In honor of Edward Keenan*, *17*, 109–130.

Hauser, M., Chomsky, N., & Fitch, W. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, *298*(5598), 1569.

Hawkins, J. A. (1994). *A performance theory of order and constituency*. Cambridge University Press.

Hawkins, J. A. (2004). *Efficiency and complexity in grammars*. Oxford University Press.

Hawkins, J. A. (2014). *Cross-linguistic variation and efficiency*. Oxford University Press.

Hays, D. G. (1960). Linguistic research at the RAND corporation. In *Proceedings of the national symposium on machine translation*. Englewood Cliffs.

Heim, I., & Kratzer, A. (1998). *Semantics in generative grammar*. Wiley-Blackwell.

Henighan, T., Carter, S., Hume, T., Elhage, N., Lasenby, R., Fort, S., Schiefer, N., & Olah, C. (2023). Superposition, memorization, and double descent. *Transformer Circuits Thread*, *6*, 24.

Hewitt, J., & Manning, C. D. (2019). A structural probe for finding syntax in word representations. In *Proceedings of the 2019 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4129–4138). Association for Computational Linguistics.

Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, *18*(7), 1527–1554.

Hofmeister, P., Jaeger, T. F., Arnon, I., Sag, I. A., & Snider, N. (2013). The source ambiguity problem: Distinguishing the effects of grammar and processing on acceptability judgments. *Language and Cognitive Processes*, *28*(1–2), 48–87.

Hornstein, N., Nunes, J., & Grohmann, K. K. (2005). *Understanding minimalism*. Cambridge University Press.

Hosseini, E., Casto, C., Zaslavsky, N., Conwell, C., Richardson, M., & Fedorenko, E. (2024a). Universality of representation in biological and

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

artificial neural networks. *bioRxiv*. https://doi.org/10.1101/2024.12.26.629294.

Hosseini, E. A., Schrimpf, M., Zhang, Y., Bowman, S., Zaslavsky, N., & Fedorenko, E. (2024b). Artificial neural network language models predict human brain responses to language even after a developmentally realistic amount of training. *Neurobiology of Language*, *5*(1), 43–63.

Hsu, A., Chater, N., & Vitányi, P. (2011). The probabilistic analysis of language acquisition: Theoretical, computational, and experimental analysis. *Cognition*, *120*(3), 380–390.

Hu, J., Gauthier, J., Qian, P., Wilcox, E., & Levy, R. (2020). A systematic assessment of syntactic generalization in neural language models. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the Association for Computational Linguistics* (pp. 1725–1744). Association for Computational Linguistics.

Hu, J., & Levy, R. (2023). Prompting is not a substitute for probability measurements in large language models. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 5040–5060). Association for Computational Linguistics.

Hu, J., Sosa, F., & Ullman, T. (2025). Re-evaluating theory of mind evaluation in large language models. *arXiv preprint arXiv:2502.21098*.

Huang, K.-J., Arehalli, S., Kugemoto, M., Muxica, C., Prasad, G., Dillon, B., & Linzen, T. (2024). Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language*, *137*, 104510.

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *Journal of Physiology*, *148*(3), 574–591.

Huben, R., Cunningham, H., Smith, L. R., Ewart, A., & Sharkey, L. (2023). Sparse autoencoders find highly interpretable features in language models. In *The twelfth international conference on learning representations*, Kigali, Rwanda.

Hubinger, E. (2019). Inductive biases stick around. AI Alignment Forum. https://www.alignmentforum.org/posts/nGqzNC6uNueum2w8T/inductive-biases-stick-around

Huh, M., Cheung, B., Wang, T., & Isola, P. (2024). The Platonic representation hypothesis. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, & F. Berkenkamp (Eds.), *Proceedings of the 41st international conference on machine learning, 235* (pp. 20617–20642). PMLR.

Hutchins, W. J. (1981). The evolution of machine translation systems. In V. Lawson (Ed.), *Translating and the computer: Practical experience of machine translation*. Aslib.

Huybregts, M. (2019). Infinite generation of language unreachable from a stepwise approach. *Frontiers in Psychology*, *10*, 425.

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, *61*(1), 23–62.

Jaeger, T. F., & Tily, H. J. (2011). On language 'utility': Processing complexity and communicative efficiency. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(3), 323–335.

Jelinek, F. (2004). Some of my best friends are linguists. Antonio Zampolli Prize presentation, 4th international conference on language resources and evaluation. http://www.lrecconf.org/lrec2004/.

Jumelet, J., Denic, M., Szymanik, J., Hupkes, D., & Steinert-Threlkeld, S. (2021). Language models use monotonicity to assess NPI licensing. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (pp. 4958–4969). Association for Computational Linguistics.

Jumelet, J., Weissweiler, L., & Bisazza, A. (2025). Multiblimp 1.0: A massively multilingual benchmark of linguistic minimal pairs. *arXiv preprint arXiv:2504.02768*.

Kallini, J., Papadimitriou, I., Futrell, R., Mahowald, K., & Potts, C. (2024). Mission: Impossible language models. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd annual meeting of the Association for Computational Linguistics (volume 1: long papers)* (pp. 14691–14714). Association for Computational Linguistics.

Keenan, E. L., & Comrie, B. (1977). Noun phrase accessibility and universal grammar. *Linguistic Inquiry*, *8*(1), 63–99.

Kim, N., & Linzen, T. (2020). COGS: A compositional generalization challenge based on semantic interpretation. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)* (pp. 9087–9105). Association for Computational Linguistics.

Kim, N., & Smolensky, P. (2021). Testing for grammatical category abstraction in neural language models. In A. Ettinger, E. Pavlick, & B. Prickett (Eds.), *Proceedings of the society for computation in linguistics 2021* (pp. 467–470). Association for Computational Linguistics.

Kim, Y., Rush, A., Yu, L., Kuncoro, A., Dyer, C., & Melis, G. (2019). Unsupervised recurrent neural network grammars. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 1105–1117). Association for Computational Linguistics.

Kirby, S., & Hurford, J. R. (2002). The emergence of linguistic structure: An overview of the iterated learning model. In *Simulating the evolution of language* (pp. 121–147). Springer.

Kluender, R., & Kutas, M. (1993). Subjacency as a processing phenomenon. *Language and Cognitive Processes*, *8*(4), 573–633.

Kobzeva, A., Arehalli, S., Linzen, T., & Kush, D. (2023). Neural networks can learn patterns of island-insensitivity in Norwegian. In *Proceedings of the society for computation in linguistics 2023* (pp. 175–185). Association for Computational Linguistics.

Kodner, J. (2022). Modeling the relationship between input distributions and learning trajectories with the Tolerance Principle. In E. Chersoni, N. Hollenstein, C. Jacobs, Y. Oseki, L. Prévot, & E. Santus (Eds.), *Proceedings of the workshop on cognitive modeling and computational linguistics* (pp. 61–67). Association for Computational Linguistics.

Kodner, J., Caplan, S., & Yang, C. (2022). Another model not for the learning of language. *Proceedings of the National Academy of Sciences*, *119*(29), e2204664119.

Kodner, J., Payne, S., & Heinz, J. (2023). Why linguistics will thrive in the 21st century: A reply to Piantadosi (2023). *arXiv preprint arXiv:2308.03228*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems, 25* (pp. 1097–1105). Curran Associates, Inc.

Kuciński, Ł., Korbak, T., Kołodziej, P., & Miłoś, P. (2021). Catalytic role of noise and necessity of inductive biases in the emergence of compositional communication. *Advances in Neural Information Processing Systems*, *34*, 23075–23088.

Kuribayashi, T., Oseki, Y., Brassard, A., & Inui, K. (2022). Context limitations make neural language models more human-like. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 10421–10436). Association for Computational Linguistics.

Kuribayashi, T., Ueda, R., Yoshida, R., Oseki, Y., Briscoe, T., & Baldwin, T. (2024). Emergent word order universals from cognitively-motivated language models. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd annual meeting of the Association for Computational Linguistics (volume 1: long papers)* (pp. 14522–14543). Association for Computational Linguistics.

Kuzborskij, I., Szepesvári, C., Rivasplata, O., Rannen-Triki, A., & Pascanu, R. (2021). On the role of optimization in double descent: A least squares study. *Advances in Neural Information Processing Systems*, *34*, 29567–29577.

Lake, B., & Baroni, M. (2018). Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In J. Dy, & A. Krause (Eds.), *Proceedings of the 35th international conference on machine learning, 80* (pp. 2873–2882). PMLR.

Lakretz, Y., Hupkes, D., Vergallito, A., Marelli, M., Baroni, M., & Dehaene, S. (2021). Mechanisms for handling nested dependencies in neural-network language models and humans. *Cognition*, *213*, 104699.

Lakretz, Y., Kruszewski, G., Desbordes, T., Hupkes, D., Dehaene, S., & Baroni, M. (2019). The emergence of number and syntax units in LSTM language models. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of*

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

the 2019 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies, volume 1 (long and short papers) (pp. 11–20). Association for Computational Linguistics.

Lampinen, A. (2024). Can language models handle recursively nested grammatical structures? A case study on comparing models and humans. *Computational Linguistics*, *50*(4), 1441–1476.

Lan, N., Chemla, E., & Katzir, R. (2024). Large language models and the argument from the poverty of the stimulus. *Linguistic Inquiry*, 1–56. https://doi.org/10.1162/ling_a_00533.

Lan, N., Geyer, M., Chemla, E., & Katzir, R. (2022). Minimum description length recurrent neural networks. *Transactions of the Association for Computational Linguistics*, *10*, 785–799.

Lappin, S., & Shieber, S. M. (2007). Machine learning theory and practice as a source of insight into universal grammar. *Journal of Linguistics*, *43*(2), 393–427.

Lasri, K., Pimentel, T., Lenci, A., Poibeau, T., & Cotterell, R. (2022). Probing for the usage of grammatical number. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th annual meeting of the Association for Computational Linguistics (volume 1: long papers)* (pp. 8818–8831). Association for Computational Linguistics.

Lau, J. H., Clark, A., & Lappin, S. (2017). Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science*, *41*(5), 1202–1241.

Lazaridou, A., Peysakhovich, A., & Baroni, M. (2017). Multi-agent cooperation and the emergence of (natural) language. In *5th international conference on learning representations, ICLR 2017*, Toulon, France.

Lederman, H., & Mahowald, K. (2024). Are language models more like libraries or like librarians? Bibliotechnism, the novel reference problem, and the attitudes of LLMs. *Transactions of the Association for Computational Linguistics*, *12*, 1087–1103.

Leong, C. S.-Y., & Linzen, T. (2023). Language models can learn exceptions to syntactic rules. In T. Hunter & B. Prickett (Eds.), *Proceedings of the society for computation in linguistics 2023* (pp. 133–144). Association for Computational Linguistics.

Lesci, P., Meister, C., Hofmann, T., Vlachos, A., & Pimentel, T. (2024). Causal estimation of memorisation profiles. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd annual meeting of the Association for Computational Linguistics (volume 1: long papers)* (pp. 15616–15635). Association for Computational Linguistics.

Levshina, N. (2022). *Communicative efficiency*. Cambridge University Press.

Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. *Advances in Neural Information Processing Systems*, *19*, 849–856.

Levy, R. P. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*(3), 375–419.

Lindsey, J., Gurnee, W., Ameisen, E., Chen, B., Pearce, A., Turner, N. L., Citro, C., Abrahams, D., Carter, S., Hosmer, B., Marcus, J., Sklar, M., Templeton, A., Bricken, T., McDougall, C., Cunningham, H., Henighan, T., Jermyn, A., Jones, A., Persic, A., Qi, Z., Thompson, T. B., Zimmerman, S., Rivoire, K., Conerly, T., Olah, C., & Batson, J. (2025). On the biology of a large language model. *Transformer Circuits Thread*.

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, *4*, 521–535.

Liu, B., Jiang, Y., Zhang, X., Liu, Q., Zhang, S., Biswas, J., & Stone, P. (2023). Llm+ p: Empowering large language models with optimal planning proficiency. *arXiv preprint arXiv:2304.11477*.

Liu, H. (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, *9*(2), 159–191.

Liu, H., Xu, C., & Liang, J. (2017). Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, *21*, 171–193.

Liu, Z., Wang, Y., Vaidya, S., Ruehle, F., Halverson, J., Soljačić, M., Hou, T. Y., & Tegmark, M. (2024). KAN: Kolmogorov–Arnold networks. *arXiv preprint arXiv:2404.19756*.

Lu, L., Xie, P., & Mortensen, D. (2024). Semi-supervised neural proto-language reconstruction. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd annual meeting of the Association for Computational Linguistics (volume 1: long papers)* (pp. 14715–14759). Association for Computational Linguistics.

Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structure. *PLOS ONE*, *5*(1), e8559.

Maass, W. (1997). Networks of spiking neurons: The third generation of neural network models. *Neural Networks*, *10*(9), 1659–1671.

Maass, W., & Markram, H. (2004). On the computational power of circuits of spiking neurons. *Journal of Computer and System Sciences*, *69*(4), 593–616.

Maddox, W. J., Benton, G., & Wilson, A. G. (2020). Rethinking parameter counting in deep models: Effective dimensionality revisited. *arXiv preprint arXiv:2003.02139*.

Madusanka, T., Pratt-Hartmann, I., & Batista-Navarro, R. (2024). Natural language satisfiability: Exploring the problem distribution and evaluating transformer-based language models. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd annual meeting of the Association for Computational Linguistics (volume 1: long papers)* (pp. 15278–15294). Association for Computational Linguistics.

Mahowald, K. (2023). A discerning several thousand judgments: GPT-3 rates the article + adjective + numeral + noun construction. In A. Vlachos, & I. Augenstein (Eds.), *Proceedings of the 17th conference of the European chapter of the Association for Computational Linguistics* (pp. 265–273). Association for Computational Linguistics.

Mahowald, K., Graff, P., Hartman, J., & Gibson, E. (2016). SNAP judgments: A small N acceptability paradigm (SNAP) for linguistic acceptability judgments. *Language*, *92*(3), 619–635.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2024). Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, *28*, 517–540.

Maloney, A., Roberts, D. A., & Sully, J. (2022). A solvable model of neural scaling laws. *arXiv preprint arXiv:2210.16859*.

Mandelkern, M., & Linzen, T. (2024). Do language models' words refer? *Computational Linguistics*, *50*(3), 1191–1200.

Manning, C. D. (1995). *Ergativity: Argument structure and grammatical relations*. Stanford University.

Manning, C. D. (2003). Probabilistic syntax. In *Probabilistic linguistics*, volume *289341*. MIT Press.

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, *117*(48), 30046–30054.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.

Mansfield, J. (2021). The word as a unit of internal predictability. *Linguistics*, *59*(6), 1427–1472.

Mansfield, J., & Kemp, C. (2023). The emergence of grammatical structure from interpredictability. *PsyArXiv*. https://doi.org/10.31234/osf.io/cjbzu.

Marjanović, S. V., Patel, A., Adlakha, V., Aghajohari, M., BehnamGhader, P., Bhatia, M., Khandelwal, A., Kraft, A., Krojer, B., Lu', X. H., Meade, N., Shin, D., Kazemnejad, A., Kamath, G., Mosbach, M., Stańczak, K., & Reddy, S. (2025). Deepseek-r1 thoughtology: Let's think about LLM reasoning. *arXiv preprint arXiv:2504.07128*.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W.H. Freeman & Company.

Martin, C. H., & Mahoney, M. W. (2021). Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *Journal of Machine Learning Research*, *22*(165), 1–73.

Martin, C. H., Peng, T., & Mahoney, M. W. (2021). Predicting trends in the quality of state-of-the-art neural networks without access to training or testing data. *Nature Communications*, *12*(1), 4122.

Marvin, R., & Linzen, T. (2018). Targeted syntactic evaluation of language models. In *Proceedings of EMNLP* (pp. 1192–1202). Association for Computational Linguistics.

McCoy, R. T., & Griffiths, T. L. (2023). Modeling rapid language learning by distilling Bayesian priors into artificial neural networks. *arXiv preprint arXiv:2305.14701*.

McCoy, R. T., Yao, S., Friedman, D., Hardy, M., & Griffiths, T. L. (2023). Embers of autoregression: Understanding large language models through the problem they are trained to solve. *arXiv preprint arXiv:2309.13638*.

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

McCoy, T., Pavlick, E., & Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th annual meeting of the Association for Computational Linguistics* (pp. 3428–3448). Association for Computational Linguistics.

McGrath, S. W., Russin, J., Pavlick, E., & Feiman, R. (2024). How can deep neural networks inform theory in psychological science? *Current Directions in Psychological Science*, *33*(5), 325–333.

Merrill, W., Petty, J., & Sabharwal, A. (2024). The illusion of state in state-space models. In *International conference on machine learning* (pp. 35492–35506). PMLR.

Merrill, W., Ramanujan, V., Goldberg, Y., Schwartz, R., & Smith, N. A. (2021). Effects of parameter norm growth during transformer training: Inductive bias from gradient descent. In M.-F. Moens, X. Huang, L. Specia, & S. W.-T. Yih (Eds.), *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 1766–1781). Association for Computational Linguistics.

Merrill, W., Sabharwal, A., & Smith, N. A. (2022). Saturated transformers are constant-depth threshold circuits. *Transactions of the Association for Computational Linguistics*, *10*, 843–856.

Meyerhoff, M. (2006). *Introducing sociolinguistics*. Routledge.

Michaelis, J. (1998). Derivational minimalism is mildly context-sensitive. In *Logical aspects of computational linguistics* (pp. 179–198). Springer.

Mikolov, T., Yih, W.-T., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies* (pp. 746–751). Association for Computational Linguistics.

Millière, R. (2024). Language models as models of language. *arXiv preprint arXiv:2408.07144*.

Millière, R., & Buckner, C. (2024). A philosophical introduction to language models-part ii: The way forward. *arXiv preprint arXiv:2405.03207*.

Misra, K., & Mahowald, K. (2024). Language models learn rare phenomena from less rare phenomena: The case of the missing aanns. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 913–929).

Mitchell, J., & Bowers, J. (2020). Priorless recurrent networks learn curiously. In D. Scott, N. Bel, & C. Zong (Eds.), *Proceedings of the 28th international conference on computational linguistics* (pp. 5147–5158). International Committee on Computational Linguistics.

Mitchell, T. M. (1980). *The need for biases in learning generalizations*. Rutgers University.

Mollica, F., & Piantadosi, S. T. (2019). Humans store about 1.5 megabytes of information during language acquisition. *Royal Society Open Science*, *6*(3), 181393.

Mordatch, I., & Abbeel, P. (2018). Emergence of grounded compositional language in multi-agent populations. In *Proceedings of the thirty-second AAAI conference on artificial intelligence and thirtieth innovative applications of artificial intelligence conference and eighth AAAI symposium on educational advances in artificial intelligence*. AAAI Press.

Moro, A., Greco, M., & Cappa, S. F. (2023). Large languages, impossible languages and human brains. *Cortex*, *167*, 82–85.

Mueller, A., Xia, Y., & Linzen, T. (2022). Causal analysis of syntactic agreement neurons in multilingual language models. In A. Fokkens & V. Srikumar (Eds.), *Proceedings of the 26th conference on computational natural language learning (CoNLL)* (pp. 95–109). Association for Computational Linguistics.

Murty, S., Sharma, P., Andreas, J., & Manning, C. (2023). Grokking of hierarchical structure in vanilla transformers. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the Association for Computational Linguistics (volume 2: short papers)* (pp. 439–448). Association for Computational Linguistics.

Musso, M., Moro, A., Glauche, V., Rijntjes, M., Reichenbach, J., Büchel, C., & Weiller, C. (2003). Broca's area and the language instinct. *Nature Neuroscience*, *6*(7), 774–781.

Nandi, A., Manning, C. D., & Murty, S. (2025). Sneaking syntax into transformer language models with tree regularization. In L. Chiruzzo, A. Ritter, & L. Wang (Eds.), *Proceedings of the 2025 conference of the nations of the Americas chapter of the Association for Computational Linguistics: Human language technologies (volume 1: long papers)* (pp. 8006–8024). Association for Computational Linguistics.

Nefdt, R. M. (2023). *Language, science, and structure: A journey into the philosophy of linguistics*. Oxford University Press.

Newmeyer, F. J. (1998). *Language form and language function*. MIT Press.

Norvig, P. (2012). Colorless green ideas learn furiously: Chomsky and the two cultures of statistical learning. *Significance*, *9*(4), 30–33.

O'Donnell, R. (2014). *Analysis of Boolean Functions*. Cambridge University Press.

Oh, B.-D., & Schuler, W. (2023). Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, *11*, 336–350.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*(6583), 607–609.

OpenAI, Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., Iftimie, A., Karpenko, A., Passos, A. T., Neitz, A., Prokofiev, A., Wei, A., Tam, A., Bennett, A., Kumar, A., Saraiva, A., Vallone, A., Duberstein, A., Kondrich, A., Mishchenko, A., Applebaum, A., Jiang, A., Nair, A., Zoph, B., Ghorbani, B., Rossen, B., Sokolowsky, B., Barak, B., McGrew, B., Minaiev, B., Hao, B., Baker, B., Houghton, B., McKinzie, B., Eastman, B., Lugaresi, C., Bassin, C., Hudson, C., Li, C. M., de Bourcy, C., Voss, C., Shen, C., Zhang, C., Koch, C., Orsinger, C., Hesse, C., Fischer, C., Chan, C., Roberts, D., Kappler, D., Levy, D., Selsam, D., Dohan, D., Farhi, D., Mely, D., Robinson, D., Tsipras, D., Li, D., Oprica, D., Freeman, E., Zhang, E., Wong, E., Proehl, E., Cheung, E., Mitchell, E., Wallace, E., Ritter, E., Mays, E., Wang, F., Such, F. P., Raso, F., Leoni, F., Tsimpourlas, F., Song, F., von Lohmann, F., Sulit, F., Salmon, G., Parascandolo, G., Chabot, G., Zhao, G., Brockman, G., Leclerc, G., Salman, H., Bao, H., Sheng, H., Andrin, H., Bagherinezhad, H., Ren, H., Lightman, H., Chung, H. W., Kivlichan, I., O'Connell, I., Osband, I., Gilaberte, I. C., Akkaya, I., Kostrikov, I., Sutskever, I., Kofman, I., Pachocki, J., Lennon, J., Wei, J., Harb, J., Twore, J., Feng, J., Yu, J., Weng, J., Tang, J., Yu, J., Candela, J. Q., Palermo, J., Parish, J., Heidecke, J., Hallman, J., Rizzo, J., Gordon, J., Uesato, J., Ward, J., Huizinga, J., Wang, J., Chen, K., Xiao, K., Singhal, K., Nguyen, K., Cobbe, K., Shi, K., Wood, K., Rimbach, K., Gu-Lemberg, K., Liu, K., Lu, K., Stone, K., Yu, K., Ahmad, L., Yang, L., Liu, L., Maksin, L., Ho, L., Fedus, L., Weng, L., Li, L., McCallum, L., Held, L., Kuhn, L., Kondraciuk, L., Kaiser, L., Metz, L., Boyd, M., Trebacz, M., Joglekar, M., Chen, M., Tintor, M., Meyer, M., Jones, M., Kaufer, M., Schwarzer, M., Shah, M., Yatbaz, M., Guan, M. Y., Xu, M., Yan, M., Glaese, M., Chen, M., Lampe, M., Malek, M., Wang, M., Fradin, M., McClay, M., Pavlov, M., Wang, M., Wang, M., Murati, M., Bavarian, M., Rohaninejad, M., McAleese, N., Chowdhury, N., Chowdhury, N., Ryder, N., Tezak, N., Brown, N., Nachum, O., Boiko, O., Murk, O., Watkins, O., Chao, P., Ashbourne, P., Izmailov, P., Zhokhov, P., Dias, R., Arora, R., Lin, R., Lopes, R. G., Gaon, R., Miyara, R., Leike, R., Hwang, R., Garg, R., Brown, R., James, R., Shu, R., Cheu, R., Greene, R., Jain, S., Altman, S., Toizer, S., Toyer, S., Miserendino, S., Agarwal, S., Hernandez, S., Baker, S., McKinney, S., Yan, S., Zhao, S., Hu, S., Santurkar, S., Chaudhuri, S. R., Zhang, S., Fu, S., Papay, S., Lin, S., Balaji, S., Sanjeev, S., Sidor, S., Broda, T., Clark, A., Wang, T., Gordon, T., Sanders, T., Patwardhan, T., Sottiaux, T., Degry, T., Dimson, T., Zheng, T., Garipov, T., Stasi, T., Bansal, T., Creech, T., Peterson, T., Eloundou, T., Qi, V., Kosaraju, V., Monaco, V., Pong, V., Fomenko, V., Zheng, W., Zhou, W., McCabe, W., Zaremba, W., Dubois, Y., Lu, Y., Chen, Y., Cha, Y., Bai, Y., He, Y., Zhang, Y., Wang, Y., Shao, Z., & Li, Z. (2024). OpenAI o1 system card.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

Papadimitriou, I., Chi, E. A., Futrell, R., & Mahowald, K. (2021). Deep subjecthood: Higher-order grammatical features in multilingual BERT. In P. Merlo, J. Tiedemann, & R. Tsarfaty (Eds.), *Proceedings of the 16th conference of the European chapter of the Association for Computational Linguistics: Main volume* (pp. 2522–2532). Association for Computational Linguistics.

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

Papadimitriou, I., & Jurafsky, D. (2020). Learning music helps you read: Using transfer to study linguistic structure in language models. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)* (pp. 6829–6839). Association for Computational Linguistics.

Papadimitriou, I., & Jurafsky, D. (2023). Injecting structural hints: Using language models to study inductive biases in language learning. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 8402–8413). Association for Computational Linguistics.

Park, Y. A., & Levy, R. (2009). Minimal-length linearizations for mildly context-sensitive dependency trees. In *Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the Association for Computational Linguistics* (pp. 335–343). Association for Computational Linguistics.

Pater, J. (2019). Generative linguistics and neural networks at 60: Foundation, friction, and fusion. *Language*, *95*(1), e41–e74.

Patil, A., Jumelet, J., Chiu, Y. Y., Lapastora, A., Shen, P., Wang, L., Willrich, C., & SteinertThrelkeld, S. (2024). Filtered corpus training (FiCT) shows that language models can generalize from indirect evidence. *Transactions of the Association for Computational Linguistics*, *12*, 1597–1615.

Payne, S. R., Kodner, J., & Yang, C. (2021). Learning morphological productivity as meaning-form mappings. *Society for Computation in Linguistics*, *4*(1), 177–187.

Pearl, L. (2022). Poverty of the stimulus without tears. *Language Learning and Development*, *18*(4), 415–454.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Empirical methods in natural language processing (EMNLP)* (pp. 1532–1543). Association for Computational Linguistics.

Pereira, F. (2000). Formal grammar and information theory: Together again? *Philosophical Transactions of the Royal Society*, *358*, 1239–1253.

Perfors, A., Tenenbaum, J. B., & Regier, T. (2013). The learnability of abstract syntactic principles. *Cognition*, *118*(3), 306–338.

Pezeshki, M., Kaba, S.-O., Bengio, Y., Courville, A., Precup, D., & Lajoie, G. (2020). Gradient starvation: A learning proclivity in neural networks. *arXiv:2011.09468 [cs, math, stat]*. arXiv: 2011.09468.

Piantadosi, S. (2023). Modern language models refute Chomsky's approach to language. *Lingbuzz Preprint, lingbuzz*, 7180.

Piantadosi, S. T., & Gallistel, C. R. (2024). Formalising the role of behaviour in neuroscience. *European Journal of Neuroscience*, *60*(5), 4756–4770.

Pimentel, T., Valvoda, J., Hall Maudslay, R., Zmigrod, R., Williams, A., & Cotterell, R. (2020). Information-theoretic probing for linguistic structure. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics* (pp. 4609–4622). Association for Computational Linguistics.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*(1–2), 73–193.

Poggio, T., Banburski, A., & Liao, Q. (2020a). Theoretical issues in deep networks. *Proceedings of the National Academy of Sciences*, *117*(48), 30039–30045.

Poggio, T., Liao, Q., & Banburski, A. (2020b). Complexity control by gradient descent in deep networks. *Nature Communications*, *11*(1), 1027.

Pollard, C., & Sag, I. A. (1994). *Head-driven phrase structure grammar*. University of Chicago Press.

Portelance, E., & Jasbi, M. (2023). The roles of neural networks in language acquisition. Retrieved from https://doi.org/10.31234/osf.io/b6978.

Potts, C. (2019). A case for deep learning in semantics: Response to pater. *Language*, *95*(1), e115–e124.

Potts, C. G. (2025). Finding linguistic structure in large language models. YouTube. YouTube video.

Pungor, J. R., Allen, V. A., Songco-Casey, J. O., & Niell, C. M. (2023). Functional organization of visual responses in the octopus optic lobe. *Current Biology*, *33*(13), 2784–2793.

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience*, *38*(33), 7255–7269.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87.

Rasin, E., Berger, I., Lan, N., Shefi, I., & Katzir, R. (2021). Approaching explanatory adequacy in phonology using Minimum Description Length. *Journal of Language Modelling*, *9*(1), 17–66.

Rathi, N., Hahn, M., & Futrell, R. (2021). An information-theoretic characterization of morphological fusion. In M.-F. Moens, X. Huang, L. Specia, & S. W.-T. Yih (Eds.), *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 10115–10120). Association for Computational Linguistics.

Rathi, N., Mehrer, J., AlKhamissi, B., Binhuraib, T., Blauch, N. M., & Schrimpf, M. (2024). TopoLM: Brain-like spatio-functional organization in a topographic language model. *arXiv preprint arXiv:2410.11516*.

Ravfogel, S., Prasad, G., Linzen, T., & Goldberg, Y. (2021). Counterfactual interventions reveal the causal effect of relative clause representations on agreement prediction. In A. Bisazza, & O. Abend (Eds.), *Proceedings of the 25th conference on computational natural language learning* (pp. 194–209). Association for Computational Linguistics.

Raviv, L., Meyer, A., & Lev-Ari, S. (2019). Larger communities create more systematic languages. *Proceedings of the Royal Society B*, *286*(1907), 20191262.

Rawski, J., & Baumont, J. (2023). Modern language models refute nothing. *Lingbuzz Preprint*. https://lingbuzz.net/lingbuzz/007203.

Ross, J. R. (1972). The category squish: Endstation Hauptwort. In P. M. Peranteau, J. N. Levi, & G. C. Phares (Eds.), *Papers from the eighth regional meeting of the Chicago Linguistic Society*. Chicago Linguistic Society.

Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel distributed processing*. MIT Press.

Rumelhart, D. E., & McClelland, J. L. (1987). Learning the past tenses of English verbs: Implicit rules or parallel distributed processing. In *Mechanisms of language acquisition* (pp. 195–248). Lawrence Erlbaum Associates, Inc.

Russin, J., McGrath, S. W., Williams, D. J., & Elber-Dorozko, L. (2024). From Frege to ChatGPT: Compositionality in language, cognition, and deep neural networks. *arXiv preprint arXiv:2405.15164*.

Ryskin, R., & Nieuwland, M. S. (2023). Prediction during language comprehension: What is next? *Trends in Cognitive Sciences*, *27*(11), 1032–1052.

Ryu, S. H., & Lewis, R. (2021). Accounting for agreement phenomena in sentence comprehension with transformer language models: Effects of similarity-based interference on surprisal and attention. In E. Chersoni, N. Hollenstein, C. Jacobs, Y. Oseki, L. Prévot, & E. Santus (Eds.), *Proceedings of the workshop on cognitive modeling and computational linguistics* (pp. 61–71). Association for Computational Linguistics.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

Sag, I. A., Wasow, T., & Bender, E. M. (2003). *Syntactic theory: A formal introduction*. Center for the Study of Language and Information Stanford.

Saldana, C., Kanampiu, P., & Culbertson, J. (2024). Gender comes first: Experimental evidence for the representation of gender closer to the noun stem than number across linguistic populations. *Linguistic Inquiry*, 1–75. https://doi.org/10.1162/ling_a_00552.

Saussure, F. (1916). *Cours de linguistique générale*. Payot, Lausanne & Paris.

Scontras, G. (2023). Adjective ordering across languages. *Annual Review of Linguistics*, *9*(1), 357–376.

Shain, C., Kean, H., Casto, C., Lipkin, B., Affourtit, J., Siegelman, M., Mollica, F., & Fedorenko, E. (2024). Distributed sensitivity to syntax and semantics throughout the language network. *Journal of Cognitive Neuroscience*, *36*(7), 1427–1471.

Shalizi, C. R., & Crutchfield, J. P. (2001). Computational mechanics: Pattern and prediction, structure and simplicity. *Journal of Statistical Physics*, *104*(3–4), 817–879.

Shibata, Y., Kida, T., Fukamachi, S., Takeda, M., Shinohara, A., Shinohara, T., & Arikawa, S. (1999). Byte pair encoding: A text compression scheme that accelerates pattern matching.

Smith, N. J., & Levy, R. P. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319.

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

Smith, N. V., Tsimpli, I.-M., & Ouhalla, J. (1993). Learning the impossible: The acquisition of possible and impossible languages by a polyglot savant. *Lingua*, *91*(4), 279–347.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, *11*(1), 1–23.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, *46*(1–2), 159–216.

Socher, R., Pennington, J., Huang, E. H., Ng, A. Y., & Manning, C. D. (2011). Semi-supervised recursive autoencoders for predicting sentiment distributions. In R. Barzilay & M. Johnson (Eds.), *Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 151–161). Association for Computational Linguistics.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In D. Yarowsky, T. Baldwin, A. Korhonen, K. Livescu, & S. Bethard (Eds.), *Proceedings of the 2013 conference on empirical methods in natural language processing* (pp. 1631–1642). Association for Computational Linguistics.

Socolof, M., Hoover, J. L., Futrell, R., Sordoni, A., & O'Donnell, T. J. (2022). Measuring morphological fusion using partial information decomposition. In N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, & S.-H. Na (Eds.), *Proceedings of the 29th international conference on computational linguistics* (pp. 44–54). International Committee on Computational Linguistics.

Someya, T., Yoshida, R., & Oseki, Y. (2024). Targeted syntactic evaluation on the Chomsky hierarchy. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)* (pp. 15595–15605). ELRA and ICCL.

Stabler, E. P. (1983). How are grammars represented? *Behavioral and Brain Sciences*, *6*(3), 391–402.

Stabler, E. P. (1997). Derivational minimalism. In *Logical aspects of computational linguistics* (pp. 68–95). Springer.

Stanojević, M., Brennan, J. R., Dunagan, D., Steedman, M., & Hale, J. T. (2023). Modeling structure-building in the brain with CCG parsing and large language models. *Cognitive Science*, *47*(7), e13312.

Staub, A. (2024). Predictability in language comprehension: Prospects and problems for surprisal. *Annual Review of Linguistics*, *11*, 17–34.

Steinert-Threlkeld, S. (2020). Toward the emergence of nontrivial compositionality. *Philosophy of Science*, *87*(5), 897–909.

Still, S. (2014). Information bottleneck approach to predictive inference. *Entropy*, *16*(2), 968–989.

Strobl, L., Merrill, W., Weiss, G., Chiang, D., & Angluin, D. (2024). What formal languages can transformers express? A survey. *Transactions of the Association for Computational Linguistics*, *12*, 543–561.

Suijkerbuijk, M., de Swart, P., & Frank, S. L. (2023). The learnability of the wh-island constraint in dutch by a long short-term memory network. In *Proceedings of the society for computation in linguistics 2023* (pp. 321–331). Association for Computational Linguistics.

Sutskever, I., Martens, J., & Hinton, G. E. (2011). Generating text with recurrent neural networks. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, Bellevue, Washington, USA, pp. 1017–1024.

Sutton, R. (2019). The bitter lesson. Available at: http://www.incompleteideas.net/IncIdeas/BitterLesson.html

Tanaka-Ishii, K. (2021). *Statistical universals of language*. Springer.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634.

Tenenbaum, J., Kemp, C., Griffiths, T., & Goodman, N. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285.

Tenney, I., Das, D., & Pavlick, E. (2019). BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th annual meeting of the Association for Computational Linguistics* (pp. 4593–4601). Association for Computational Linguistics.

Timkey, W., & Linzen, T. (2023). A language model with limited memory capacity captures interference in human sentence processing. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 8705–8720). Association for Computational Linguistics.

Tishby, N., Pereira, F. C., & Bialek, W. (1999). The information bottleneck method. In *Proceedings of the 37th annual Allerton conference on communication, control, and computing*, Monticello, IL, USA, pp. 368–377. https://arxiv.org/abs/physics/0004057.

Tishby, N., & Zaslavsky, N. (2015). Deep learning and the information bottleneck principle. In *2015 IEEE information theory workshop (ITW)* (pp. 1–5). IEEE.

Tseng, Y.-H., Shih, C.-F., Chen, P.-E., Chou, H.-Y., Ku, M.-C., & Hsieh, S.-K. (2022). CxLM: A construction and context-aware language model. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the thirteenth language resources and evaluation conference* (pp. 6361–6369). European Language Resources Association.

Upadhye, S., & Futrell, R. (2025). Examining future context predictability effects in word-form variation and word choice. In *Proceedings of the Cognitive Science Society*.

Üstün, A., Aryabumi, V., Yong, Z., Ko, W.-Y., D'souza, D., Onilude, G., Bhandari, N., Singh, S., Ooi, H.-L., Kayid, A., Vargus, F., Blunsom, P., Longpre, S., Muennighoff, N., Fadaee, M., Kreutzer, J., & Hooker, S. (2024). Aya model: An instruction finetuned open-access multilingual language model. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd annual meeting of the Association for Computational Linguistics (volume 1: Long papers)* (pp. 15894–15939). Association for Computational Linguistics.

Valle-Perez, G., Camargo, C. Q., & Louis, A. A. (2018). Deep learning generalizes because the parameter-function map is biased towards simple functions. *arXiv preprint arXiv:1805.08522*.

van Schijndel, M., & Linzen, T. (2018). Modeling garden path effects without explicit hierarchical syntax. In *Proceedings of the 40th annual meeting of the cognitive science society* (pp. 2603–2608). The Cognitive Science Society.

van Schijndel, M., & Linzen, T. (2021). Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive Science*, *45*(6), e12988.

Vázquez Martínez, H., Lea Heuser, A., Yang, C., & Kodner, J. (2023). Evaluating neural language models as cognitive models of language acquisition. In D. Hupkes, V. Dankers, K. Batsuren, K. Sinha, A. Kazemnejad, C. Christodoulopoulos, R. Cotterell, & E. Bruni (Eds.), *Proceedings of the 1st GenBench workshop on (Benchmarking) generalisation in NLP* (pp. 48–64). Association for Computational Linguistics.

Vijay-Shanker, K., Weir, D. J., & Joshi, A. K. (1987). Characterizing structural descriptions produced by various grammatical formalisms. In *Proceedings of the 25th annual meeting of the Association for Computational Linguistics* (pp. 104–111). Association for Computational Linguistics.

Voita, E., & Titov, I. (2020). Information-theoretic probing with minimum description length. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)* (pp. 183–196). Association for Computational Linguistics.

Wang, S. I., & Manning, C. D. (2012). Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th annual meeting of the Association for Computational Linguistics (volume 2: short papers)* (pp. 90–94). Association for Computational Linguistics.

Wang, W., Vong, W. K., Kim, N., & Lake, B. M. (2023). Finding structure in one child's linguistic experience. *Cognitive Science*, *47*(6), e13305.

Warstadt, A., & Bowman, S. R. (2022). What artificial neural networks can teach us about human language acquisition. In S. Lappin, & J.-P. Bernardy (Eds.), *Algebraic structures in natural language*. Taylor and Francis.

Warstadt, A., Mueller, A., Choshen, L., Wilcox, E., Zhuang, C., Ciro, J., Mosquera, R., Paranjabe, B., Williams, A., Linzen, T., et al. (2023). Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM challenge at the 27th conference on computational natural language learning*. Association for Computational Linguistics.

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., & Bowman, S. R. (2020). BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, *8*, 377–392.

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

Weir, D. J. (1988). *Characterizing mildly context-sensitive grammar formalisms*. University of Pennsylvania.
Weissweiler, L., He, T., Otani, N., Mortensen, D. R., Levin, L., & Schütze, H. (2023). Construction grammar provides unique insight into neural language models. In C. Bonial & H. Tayyar Madabushi (Eds.), *Proceedings of the first international workshop on construction grammars and NLP (CxGs+NLP, GURT/SyntaxFest 2023)* (pp. 85–95). Association for Computational Linguistics.
Weissweiler, L., Hofmann, V., Köksal, A., & Schütze, H. (2022). The better your syntax, the better your semantics? Probing pretrained language models for the English comparative correlative. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 10859–10882). Association for Computational Linguistics.
Wheeler, J. A. (1989). Information, physics, quantum: The search for links. In *Proceedings of 3rd International Symposium on the Foundations of Quantum Mechanics*, pp. 354–368, Tokyo.
White, J. C., Pimentel, T., Saphra, N., & Cotterell, R. (2021). A non-linear structural probe. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, & Y. Zhou (Eds.), *Proceedings of the 2021 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies* (pp. 132–138). Association for Computational Linguistics.
Wilcox, E., Levy, R., & Futrell, R. (2019a). Hierarchical representation in neural language models: Suppression and recovery of expectations. In T. Linzen, G. Chrupała, Y. Belinkov, & D. Hupkes (Eds.), *Proceedings of the 2019 ACL workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP* (pp. 181–190). Association for Computational Linguistics.
Wilcox, E., Levy, R., Morita, T., & Futrell, R. (2018). What do RNN language models learn about filler–gap dependencies? In *Proceedings of BlackboxNLP*. Association for Computational Linguistics
Wilcox, E., Qian, P., Futrell, R., Ballesteros, M., & Levy, R. (2019b). Structural supervision improves learning of non-local grammatical dependencies. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 3302–3312). Association for Computational Linguistics.
Wilcox, E. G., Futrell, R., & Levy, R. (2023a). Using computational models to test syntactic learnability. *Linguistic Inquiry*, *55*, 1–44.
Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., & Levy, R. (2020). On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the 42nd annual meeting of the cognitive science society* (pp. 1707–1713). Cognitive Science Society.
Wilcox, E. G., Pimentel, T., Meister, C., Cotterell, R., & Levy, R. P. (2023b). Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, *11*, 1451–1470.
Wilson, A. G. (2025). Deep learning is not so mysterious or different. *arXiv preprint arXiv:2503.02113*.
Winograd, T. (1972). Understanding natural language. *Cognitive Psychology*, *3*(1), 1–191.
Wolfram, S. (2023). *What is ChatGPT doing: . . . and why does it work?* Wolfram Media.
Wolpert, D. H., Macready, W. G., et al. (1995). *No free lunch theorems for search*. Citeseer.
Xu, T., Kuribayashi, T., Oseki, Y., Cotterell, R., & Warstadt, A. (2025). Can language models learn typologically implausible languages? *arXiv preprint arXiv:2502.12317*.
Xu, W., Chon, J., Liu, T., & Futrell, R. (2023). The linearity of the effect of surprisal on reading times across languages. In *Findings of the association for computational linguistics: EMNLP 2023* (pp. 15711–15721). Association for Computational Linguistics.
Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624.
Yang, X., Aoyama, T., Yao, Y., & Wilcox, E. (2025). Anything goes? A cross-linguistic study of (im) possible language learning in LMs. *arXiv preprint arXiv:2502.18795*.
Yang, Y., & Piantadosi, S. T. (2022). One model for the learning of language. *Proceedings of the National Academy of Sciences*, *119*(5), e2021865119.
Yao, Q., Misra, K., Weissweiler, L., & Mahowald, K. (2025). Both direct and indirect evidence contribute to dative alternation preferences in language models. *arXiv preprint arXiv:2503.20850*.
Yedetore, A., Linzen, T., Frank, R., & McCoy, R. T. (2023). How poor is the stimulus? Evaluating hierarchical generalization in neural networks trained on child-directed speech. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the Association for Computational Linguistics (volume 1: Long papers)* (pp. 9370–9393). Association for Computational Linguistics.
Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, *64*(3), 107–115.
Zhao, J., Gao, R., & Brennan, J. R. (2025). Decoding the neural dynamics of headed syntactic structure building. *Journal of Neuroscience*, *45*(17), e2126242025.
Ziv, I., Lan, N., Chemla, E., & Katzir, R. (2025). Large language models as proxies for theories of human linguistic cognition. *arXiv preprint arXiv:2502.07687*.

Open Peer Commentary

Mutually assured disruption: why LLMs change the game for child language acquisition research, and vice versa

Ben Ambridge*

University of Manchester, UK
ben.ambridge@manchester.ac.uk

*Corresponding author.

doi:10.1017/S0140525X26104592, e199

Abstract

LLMs change the game for child language acquisition research, because they constitute an existence proof that – counter to most prevailing theoretical accounts – it is possible for learners to acquire a productive system of grammar solely from exposure to individual utterances, without the need for abstract grammatical rules or categories. Conversely, LLM development can benefit from insights from child language research.

Let me quickly contextualise this commentary by saying I'm no LLM booster. On the contrary, I see LLMs as a disaster for both our intellectual environment (e.g., Guest et al., 2025) and our environment (e.g., Luccioni, Viguier & Ligozat, 2023). Either AI fulfils its promise, and we're all out of a job, or it doesn't, and we'll have burned academia, the global economy and a tonne of fossil fuels for nothing. If I could wave a magic wand and uninvent LLMs, I would.

But while LLMs are (still?) here, we may as well learn from them. And even as a committed non-fan, it seems to me that LLMs

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

have already changed the game for child language acquisition research.

How so? Even today, my impression of the field is that most child language acquisition researchers are essentially essentialists. They see the child's task as – in broad brush terms – acquiring abstract categories like VERB, NOUN, and DETMINER, and rules for combining them into sentences. This assumption is, of course, explicit in Chomskyan accounts of language development, which assume that these abstract categories are acquired very early, or (whisper it) perhaps even innate. But it is surprisingly common to see constructivist accounts of how particular abstract categories or constructions are "acquired," and surprisingly rare to see ones that eschew these types of freestanding essentialist categories altogether.

Why? Humans are natural essentialists (e.g., Gelman, 2004; Hull, 1976; Mayr, 1982). In a confusing world, it's comforting to think in terms of categories with sharp boundaries – black or white; lion or tiger; good or evil – and when it turns out, as it always does, that "it's actually a lot more complicated than that," lots of people get very upset. Linguists and language acquisition researchers are by no means immune (e.g., Croft, in press). It would be nice if we could understand language development as the acquisition of simple rules for manipulating discrete categories. And, as usual, attempts to say "it's actually a lot more complicated than that" – i.e., it's some kind of very complex functionally-based, input-based analogical generalisation – have historically been waved away (Adger, 2020; Goodluck, 2020; Hartshorne, 2020; Koring et al., 2020; Pearl, 2014; Pinker & Ullman, 2002), or even mocked (e.g., Pinker, 1979):

> For example, sentences 3(a-d)
>
> 3. (a) Hottentots must survive.
> (b) Hottentots must fish.
> (c) Hottentots eat fish.
> (d) Hottentots eat rabbits.
>
> would seduce a distributional analysis learner into . . . the production of "Hottentots must rabbits", "Hottentots eat survive", and other monstrosities.

"Come on then, " these papers said, "If it's possible to learn grammar without rules and categories, show me a system that does so." LLMs change the game for child language acquisition, then, because they do just this. They're an existence proof that a system that doesn't start out with abstract grammatical rules and categories – or even build them, except in the most distributed, emergent sense – can acquire abstract grammar. Yes, LLMs have all kinds of problems – they don't have meanings grounded in the real world, they require implausible amounts of input, etc. (e.g., Frank, 2023); but that doesn't matter: they're an *existence proof*. As many of us have been saying all along (e.g., Ambridge & Lieven, 2015; Ambridge, 2020a,b), you *can* learn grammar without abstract rules and categories; the question is no longer "whether" but "how."

So that's how LLMs change the game for child language acquisition research; what about the "vice versa"? This, I must admit, is more a pipe dream than a reality. But is it too vain to hope that we might be able to reduce both AI slop and its environmental impact by building into LLM-development insights from language-development research?

Children's language learning is multimodal and embodied, containing rich real-world semantic representations (e.g., Reggin et al., 2023). Communicative function is key: children use language to get across *what they want to say* (e.g., Bannard, Rosner & Matthews, 2017). Children hear input forms not at random, but in bursts of identical/similar forms (e.g., Cychosz et al., 2025). They are subject to developmentally decreasing memory and processing limitations that narrow the "context window" (Gathercole et al., 2004; Pine, Freudenthal, & Gobet, 2023), and result in non-linearities in learning (e.g., Brooks & Meltzoff, 2008). They are overly sensitive to high-frequency chunks, in ways that yield characteristic errors (Ambridge et al., 2015). They learn, at least at first, not from tokenised text but from sounds better approximated as vectors of phonological features (Albright & Hayes, 2003; Rubehn et al., 2024).

For these reasons, LLMs ignore developmental child language research at their peril, and vice versa. Disruption is mutually assured.

Acknowledgements. None.

Financial support. None.

Competing interests. None.

References

Adger, D. (2020). Syntax and the failure of analogical generalization: A commentary on Ambridge (2020). *First Language*, *40*(5–6), 560–563.

Albright, A., & Hayes, B. (2003). Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition*, *90*(2), 119–161.

Ambridge, B. (2020a). Against stored abstractions: A radical exemplar model of language acquisition. *First Language*, *40*(5–6), 509–559 https://doi.org/10.1177/0142723719869731

Ambridge, B. (2020b). Abstractions made of exemplars or you're all right and I've changed my mind. Response to commentators. *First Language*, *40*(5–6), 640–659 https://doi.org/10.1177/0142723720949723

Ambridge, B., & Lieven, E. V. M. (2015). A constructivist account of child language acquisition. In B. MacWhinney, & W. O'Grady (Eds.), *Handbook of language emergence* (pp. 478–510). Wiley Blackwell.

Ambridge, B., Rowland, C. F., Theakston, A. L. & Kidd, E. J. (2015). The ubiquity of frequency effects in first language acquisition. *Journal of Child Language*, *42*(2), 239–273.

Bannard, C., Rosner, M., & Matthews, D. (2017). What's worth talking about? Information theory reveals how children balance informativeness and ease of production. *Psychological Science*, *28*(7), 954–966.

Brooks, R., & Meltzoff, A. N. (2008). Infant gaze following and pointing predict accelerated vocabulary growth through two years of age: A longitudinal, growth curve modeling study. *Journal of Child Language*, *35*(1), 207–220.

Croft, W., (in press). Lineage-based linguistics. In W. Arka, D. Barth, & H. Bergqvist (Eds.), *Linguistic worlds and social minds: What language documentation and linguistic diversity can tell us about the nature and origin of language*. Oxford University Press. https://www.unm.edu/~wcroft/Papers/Lineage-based-final.pdf

Cychosz, M., Romeo, R. R., Edwards, J. R., & Newman, R. S. (2025). Bursty, irregular speech input to children predicts vocabulary size. *Developmental Science*, *28*(1), e13590.

Frank, M. C. (2023). Bridging the data gap between children and large language models. *Trends in Cognitive Sciences*, *27*(11), 990–992.

Gathercole, S. E., Pickering, S. J, Ambridge, B. & Waring, H. (2004). The structure of working memory from 4-15 years of age. *Developmental Psychology*, *40*(2), 177–190.

Gelman, S. A. (2004). Psychological essentialism in children. *Trends in Cognitive Sciences*, *8*(9), 404–409.

Goodluck, H. (2020). *Language acquisition by children: A linguistic introduction*. Edinburgh University Press.

Guest, O., Suarez, M., Müller, B. C. N., van Meerkerk, E., Beverborg, A. O. G., de Haan, R., Reyes Elizondo, A., Blokpoel, M., Scharfenberg, N., Kleinherenbrink, A., Camerino, I., Woensdregt, M., Monett, D., Brown, J., Avraamidou, L., Alenda-Demoutiez, J., Hermans, F., & van Rooij, I. (2025).

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

Against the uncritical adoption of 'AI' technologies in academia. Zenodo. https://doi.org/10.5281/zenodo.17065099
Hartshorne, J. K. (2020). The many blessings of abstraction: A commentary on Ambridge (2020). *First Language*, *40*(5–6), 581–584.
Hull, D. L. (1976). Are species really individuals? *Systematic Zoology*, *25*, 174–191.
Koring, L., Giblin, I., Thornton, R., & Crain, S. (2020). Like dishwashing detergents, all analogies are not the same: A commentary on Ambridge (2020). *First Language*, *40*(5–6), 596–599.
Luccioni, A. S., Viguier, S., & Ligozat, A. L. (2023). Estimating the carbon footprint of bloom, a 176b parameter language model. *Journal of Machine Learning Research*, *24*(253), 1–15.
Mayr, E. (1982). *The Growth of biological thought: Diversity, evolution, inheritance*. Belknap Press.
Pearl, L. (2014). Evaluating learning-strategy components: Being fair (Commentary on Ambridge, Pine, and Lieven). *Language*, *90*(3), e107–e114.
Pine, J.M., Freudenthal, D. & Gobet, F. (2023). Building a unified model of the optional infinitive stage: Simulating the cross-linguistic pattern of verb-marking error in typically developing children and children with developmental language disorder. *Journal of Child Language*, *50*(6), 1336–1352.
Pinker, S. (1979). Formal models of language learning. *Cognition*, *7*(3), 217–283.
Pinker, S., & Ullman, M. T. (2002). The past and future of the past tense. *Trends in cognitive sciences*, *6*(11), 456–463.
Reggin, L. D., Franco, L. E. G., Horchak, O. V., Labrecque, D., Lana, N., Rio, L., & Viglocco, G. (2023). Consensus paper: Situated and embodied language acquisition. *Journal of Cognition*, *6*(1), 63.
Rubehn, A., Nieder, J., Forkel, R. & List, J. M. (2024). Generating feature vectors from phonetic transcriptions in cross-linguistic data formats. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2024,* 205–216. Irvine, CA.

Linguistics should not stop worrying about languages other than English: A commentary on Futrell and Mahowald

Bjarki Ármannsson[a,b]*, Iris Edda Nowenstein[b] and Einar Freyr Sigurðsson[a]

[a]The Árni Magnússon Institute for Icelandic Studies, Iceland, and [b]University of Iceland, Iceland
bjarki.armannsson@arnastofnun.is
einar.freyr.sigurdsson@arnastofnun.is
irisen@hi.is

*Corresponding author.

doi:10.1017/S0140525X26104658, e200

Abstract

Futrell and Mahowald's assumptions about parallels between language models and human language are primarily based on LMs' performance in English. In our commentary, we discuss how this impacts some of their key takeaways and highlight LM shortcomings in languages other than English (specifically, Icelandic), which are only alluded to in the target article.

Although we agree with Futrell and Mahowald's (F&M) overall stance as we understand it – that language models (LMs) can serve as tools for investigating features of human language and that theoretical and computational linguistics can inform each other going forward – we question the validity and generalizability of some of F&M's conclusions cross-linguistically. Just as the language sciences (not least early work in generative grammar) have been criticized for basing assumptions primarily on examples from English (Blasi et al., 2022; Evans & Levinson, 2009), some of F&M's key takeaways (notably, that LMs' success "strengthens the case for existing usage-based language theories") are almost exclusively based on research on model capabilities in English, impacting the validity of LMs as a language-agnostic model of human language. This is (briefly) recognized as a "caveat" at the outset, but no effort is made to engage with the obvious argument that spending years building statistical models that optimize primarily over English text data can easily lead to overfitting to English linguistic structure (an argument made at least as early as Bender (2011) about machine-learning systems in general). This argument is not only theoretical; previous work has produced examples where LMs show a clear preference for English-like grammars over non-English-like ones (Davis and van Schijndel 2020; Davis 2022; Dyer, Melis, & Blunsom 2019). Most egregiously, in section 4.4, the terms "real English text" and "human language" are practically treated as equivalent when discussing the learning of "impossible languages."

This clearly impacts the discussion of whether LMs can be considered suitable models of human language. While F&M discuss how autoregressive LMs "show a bias towards information locality" obtained through their next-word prediction training setup and how this bias "helps language learning and is likely shared with humans," it is by no means clear that all languages are equally well-suited to the next-word prediction objective. Indeed, recent work shows that certain very robust features of Icelandic are not effectively learned by LMs trained on more Icelandic data than a human is exposed to during the language acquisition phase, likely by orders of magnitude.

In Ármannsson et al. (2026), we find that state-of-the-art LMs (as of 2024) fall remarkably short of human accuracy for robust noun-phrase internal predicate-agreement patterns in Icelandic when rating clearly grammatical/ungrammatical sentences, often preferring ungrammatical gender agreement with a pre-adjectival subject in the presence of a post-adjectival predicate. While native speakers showed 99% acceptance for the grammatical sentences and 0.3% for ungrammatical ones, the best-performing model recorded 57% accuracy, and almost all fell below chance (Figure 1). In addition, assuming that children are exposed to roughly seven to ten million words per year (Gilkerson et al., 2017), this agreement pattern seems robust in children "trained" on no more than twenty million words, presumably a fraction of the models' Icelandic training data (the largest corpus of Icelandic available online, the Icelandic Gigaword Corpus (IGC; Barkarson, Steingrímsson, & Hafsteinsdóttir, 2022), contains approximately 2.6 billion words).

While the suitability of LMs as models of human language obviously does not hinge on this or any other single experiment, it is worth contemplating that F&M, and much of the literature they refer to, focus on work that shows LMs successfully modelling – to various degrees – "nontrivial formal linguistic patterns" in English (often quite rare constructions, e.g., Misra & Mahowald, 2024). Meanwhile, examples like ours show massive models failing to successfully generalize a ubiquitous pattern that appears to be both *categorical* in native speakers and *unanimous* in corpora of Icelandic (we found zero counterexamples in our search of the morphologically tagged IGC). This is in direct contrast with, e.g., the commonly cited work of Linzen, Dupoux, and Goldberg (2016), who probe models through an English pattern that is known to be

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

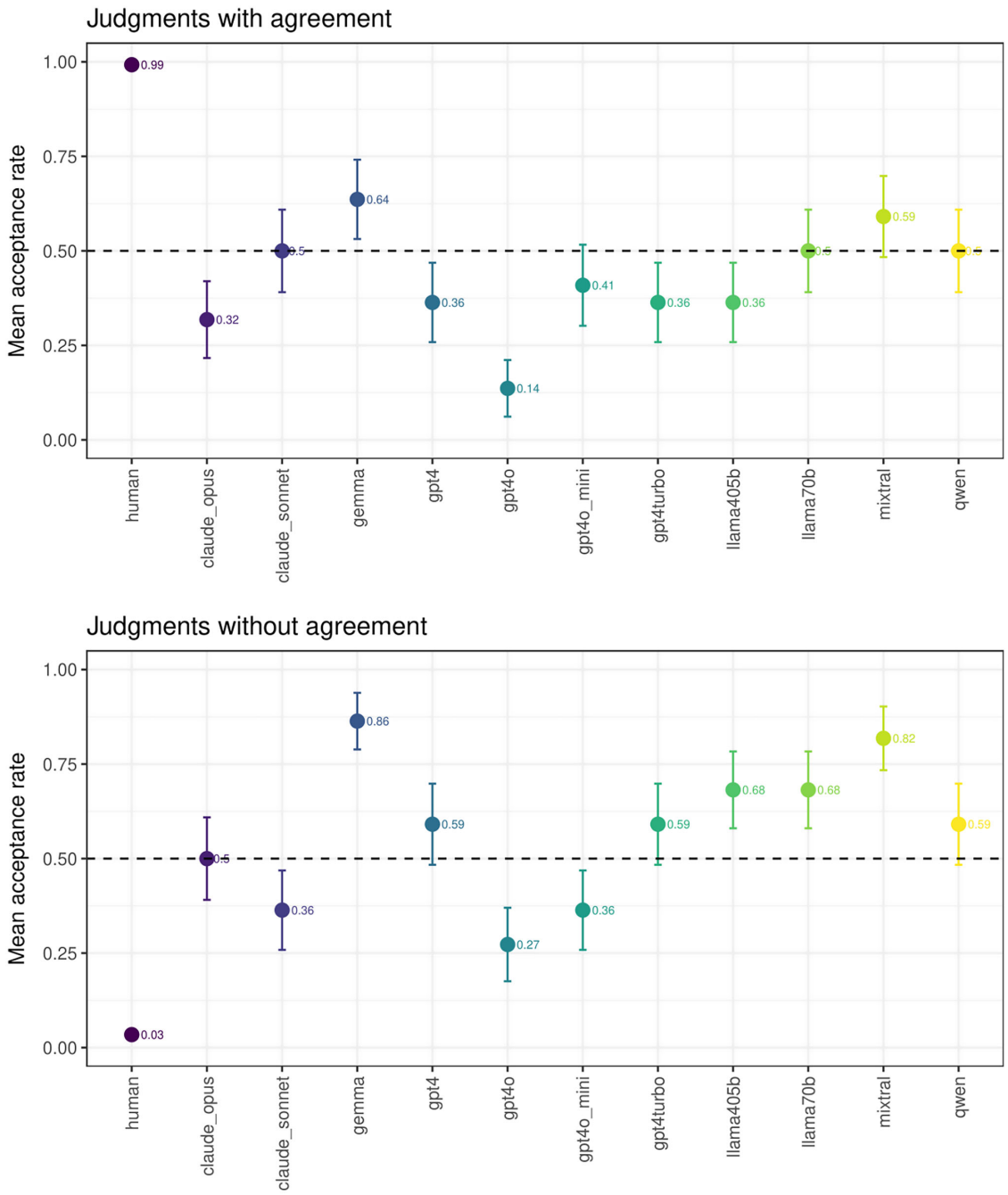


Figure 1 (Ármannsson et al.). The mean acceptance rate of sentences in the gender agreement grammaticality judgment tasks in Ármannsson et al. (2026), shown for only those sentences that contained noun-phrase-internal predicate agreement (above) and those that did not (below). Native speakers categorically accept sentences containing noun-phrase internal predicate agreement and reject those without, while no model reliably matches the judgments of human participants.

variable in speakers (or in F&M's framing, "human accuracy in producing the right verb forms [. . .] is 85%") and dependent on various syntactic and semantic factors (den Dikken, 2001).

Overall, F&M make an obvious and commendable effort to balance their arguments with the necessary caveats, which is why we find it surprising that they eventually arrive at the conclusion that "LMs are a proof of concept for gradient, usage-based theories of language." Usage-based theories of language, as all theories of language, are meant to be language agnostic. (They are also meant to account for how language is learned with the time and input available to a child learner, which remains the most urgent obstacle to treating LMs as theories of language. In our view, this significant issue is not addressed well enough by F&M, but we leave this line of argument to other commentators.) There is a major leap between the well-argued claim that LMs "have learned nontrivial formal linguistic patterns better than many expected was possible" and the conclusion that their success "strengthens the case for existing usage-based language theories based on gradient representations." In order to make this leap, F&M would need to present considerably more cross-linguistic evidence than they do and extend their

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

argument against modern computational and neurolinguistic work rooted in generative linguistics (e.g., Belth et al., 2021; Gwilliams et al., 2025; Payne, Kodner, & Yang 2021). This work is cited to some extent, but not included under the umbrella of generative linguistics, perhaps because it does not neatly fit into a more simplistic binary characterization pitting a usage-based/statistical tradition against a generative/"anti-statistical" one.

LMs' successful performance in processing and outputting English text can lead to overhyped claims about their approximation of the human language acquisition process in general, across any language. Similarly, it can lead to claims that LMs provide us with clear answers in debates opposing usage-based and generative theories of language, despite a growing body of work showing parallels between various aspects of linguistic theory, including generative ones, and LM performance (e.g., Caucheteux & King, 2022; Desbordes et al., 2023; Simon et al., 2024). Hopefully, future work will focus more on the nuances present in F&M's target article and less on reductive claims about LM performance and its impact on modern linguistic theory.

Acknowledgements. None.

Financial support. The original research described in this commentary was partly funded by the Icelandic Research Fund's Doctoral Student Grant no. 2510917-051 (Bjarki Ármannsson) and by the Icelandic government's Language Technology Program.

Competing interests. The authors declare no competing interests.

References

Ármannsson, B., Ingimundarson, F. Á., Nowenstein, I. E., & Sigurðsson, E. F. (2026). Icelandic Gender Agreement Shows LLMs' Failure to Effectively Generalise. To appear in *University of Pennsylvania Working Papers in Linguistics* 32(1).

Barkarson, S., Steingrímsson, S., & Hafsteinsdóttir, H. (2022). Evolving large text corpora: Four versions of the Icelandic Gigaword Corpus. *Proceedings of the Thirteenth Language Resources and Evaluation Conference,* 2371–2381. European Language Resources Association. https://aclanthology.org/2022.lrec-1.254/

Belth, C., Payne, S., Beser, D., Kodner, J., & Yang, C. (2021). The greedy and recursive search for morphological productivity. *The 43rd Annual Meeting of the Cognitive Science Society (CogSci),* 2869–2875.

Bender, E. M. (2011). On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology*, *6*(3), 1–26. https://doi.org/10.33011/lilt.v6i.1239

Blasi, D. E., Henrich, J., Adamou, E., Kemmerer, D., & Majid, A. (2022). Over-reliance on English hinders cognitive science. *Trends in Cognitive Sciences*, *26*(12), 1153–1170. https://doi.org/10.1016/j.tics.2022.09.015

Caucheteux, C., & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Communications Biology*, *5*, 134. https://doi.org/10.1038/s42003-022-03036-1

Davis, F. L. (2022). *On the limitations of data: Mismatches between neural models of language and humans.* Unpublished doctoral dissertation, Cornell University.

Davis, F., & van Schijndel, M. (2020). Recurrent neural network language models always learn English-like relative clause attachment. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1979–1990. Association for Computational Linguistics. https://aclanthology.org/2020.acl-main.179/

den Dikken, M. (2001). "Pluringulars", pronouns and quirky agreement. *The Linguistic Review*, *18*, 19–41. https://doi.org/10.1515/tlir.18.1.19

Desbordes, T., Lakretz, Y., Chanoine, V., Oquab, M., Badier, J.-M., Trébuchon, A., Carron, R., Bénar, C.-G., Dehaene, S., & King, J.-R. (2023). Dimensionality and ramping: Signatures of sentence integration in the dynamics of brains and deep language models. *Journal of Neuroscience*, *43*(29), 5350–5364. https://doi.org/10.1523/JNEUROSCI.1163-22.2023

Dyer, C., Melis, G., & Blunsom, P. (2019). A critical analysis of biased parsers in unsupervised parsing. arXiv:1909.09428v. https://arxiv.org/pdf/1909.09428

Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, *32*(5), 429–448. https://doi.org/10.1017/S0140525X0999094X

Gilkerson, J., Richards, J. A., Warren, S. F., Montgomery, J. K., Greenwood, C. R., Kimbrough Oller, D., Hansen, J. H. L., & Paul, T. D. (2017). Mapping the early language environment using all-day recordings and automated analysis. *American Journal of Speech-Language Pathology*, *26*(2), 248–265. https://doi.org/10.1044/2016_AJSLP-15-0169

Gwilliams, L., Marantz, A., Poeppel, D., & King, J.-R. (2025). Hierarchical dynamic coding coordinates speech comprehension in the human brain. *Proceedings of the National Academy of Sciences of the United States of America*, *122*(42), e2422097122. https://doi.org/10.1073/pnas.2422097122

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics 4*, 521–535. https://aclanthology.org/Q16-1037/

Misra, K., & Mahowald, K. (2024). Language models learn rare phenomena from less rare phenomena: The case of the missing AANNs. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 913–929. Association for Computational Linguistics. https://aclanthology.org/2024.emnlp-main.53/

Payne, S., Kodner, J., & Yang, C. (2021). Learning morphological productivity as meaning-form mappings. *Proceedings of the Society for Computation in Linguistics 2021*, 177–187. Association for Computational Linguistics. https://aclanthology.org/2021.scil-1.17/

Simon, P. J. D., d'Ascoli, S., Chemla, E., Lakretz, Y., & King, J.-R. (2024). A Polar coordinate system represents syntax in large language models. *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=x2780VcMOI

Large language models are not about natural language

Johan J. Bolhuis[a]*, Andrea Moro[b], Stephen Crain[c] and Sandiway Fong[d]

[a]Department of Psychology, University of Cambridge, Cambridge, UK; [b]University School for Advanced Studies, Pavia & Scuola Normale Superiore, Pisa, Italy; [c]Department of Linguistics, Macquarie University, Sydney, Australia and [d]Department of Linguistics, University of Arizona, Tucson, Arizona, USA
j.j.bolhuis@uu.nl
andrea.moro@iusspavia.it
stephen.crain@mq.edu.au
sandiway@arizona.edu

*Corresponding author.

doi:10.1017/S0140525X26104622, e201

Abstract

Large Language Models are useless for linguistics, as they are probabilistic models that require a vast amount of data to analyze externalized strings of words. In contrast, human language is underpinned by a mind-internal computational system that recursively generates hierarchical thought structures. The language system grows with minimal external input and can readily distinguish between real language and impossible languages.

The title of the target article is a little confusing, particularly with regard to what is meant by "language models." In any science, the starting point for research is the formulation of a model, a theory of the underlying mechanisms that could go towards explaining the particular phenomenon at hand. In linguistics, the Strong Minimalist Thesis (Chomsky, 2000) is such a theory, a model for the mechanism of language. It becomes clear from the target article itself that Futrell and Mahowald specifically refer to Large Language Models (LLMs). We disagree with all the arguments that the authors offer to support their claim that LLMs can be useful for our understanding of "linguistic structure, language processing, and learning." This is because human language is a computational system that works in a fundamentally different way from LLMs (Bolhuis et al., 2023, 2024; Fong, 2025; Moro, Greco, & Cappa, 2023). LLMs are probabilistic models that analyze externalized – "flattened" – strings of words. As such, they are not new: already in 1913, Markov (1913/2006) published a probabilistic analysis of vowels and consonants confirming a strong bias in favor of an alternating vowel/consonant pattern in the poetry of Pushkin's *Eugene Onegin*. The probabilistic nature of LLMs is in complete contrast to the recursive function by which the human mind generatively forms hierarchical thought structures that may or may not be externalizable (Everaert et al., 2015; Fong, 2025; Friederici et al., 2017). The nature of those structures determines meaning (semantics) (Everaert et al., 2015). LLMs are not generative in this sense, and they do not form structures that determine meaning.

In addition, LLMs "acquire" language in a fundamentally different way from human infants, who can build syntactic structures in their mind even without any relevant input (Bolhuis et al., 2023; Crain, Koring, & Thornton, 2017; Yang et al., 2017) – the "Poverty of the Stimulus" argument that the target authors discuss. This is in stark contrast to the enormous amount of training plus trillion-parameter models that are absolutely required for LLMs to generate natural output (Fong, 2025). The contrast between the two is aptly illustrated by the differences in energy use. Although it is not often made public, we know that the environmental impact and power requirements of the latest language models are quite staggering. For example, it has been reported that xAI in Memphis, TN requires 70MW to run 100,000 GPUs concurrently, a power that the local utility was initially unable to supply. As a result, 18 natural gas generators (2.5MW each) had to be parked outside (Kerr, 2024). Similarly, Google has ordered six or seven small nuclear reactors for its AI datacenters, according to Lawson (2024). In contrast, it is estimated that the human brain consumes about 20W, not all of which is available to language (Ling, 2001).

The manner in which LLMs learn linguistic patterns is also quite unlike the developmental stages children go through in the transition to adult language. At some stages, children's production and understanding of language systematically differ from that of adults, and include linguistic structures that are not attested in the adult input (Crain, 2012). The stages of child language development have no semblance to the way LLMs assimilate linguistic patterns from training regimes. LLMs therefore, shed no light on the cognitive architecture that enables human children to acquire languages.

Unlike human brains, machines based on LLMs are unable to distinguish between possible and impossible languages (Bolhuis et al., 2024; Moro, 2016; Moro et al., 2023). That the human brain can make this distinction is based on convergent experiments, exploiting both actual languages (Musso et al., 2003) and invented languages (Tettamanti et al., 2002). The common paradigm consists of contrasting rules based on linear sequences of words vs. hierarchical constituents, which are the ones based on all and only human languages (Rizzi, 2009), showing a selective inhibition of a network involving Broca's area for impossible languages only. The target authors address this issue and argue that LLMs have "inductive biases [that] meaningfully align with linguistic structure." They specifically refer to a study of their own (Kallini et al., 2024) that purports to show that LLMs learn from English text significantly better than from impossible variants. However, the authors fail to mention other work that has shown that their conclusions are unwarranted. As Bowers (2025) pointed out, the worst performance of the LLMs in Kallini et al. (2024) was with a "language" where the words were shuffled randomly within sentences. As Bowers (2025) rightly concludes, this has no structure at all, and thus is not a language, impossible or otherwise. The other examples with structured languages resulted in minimal or no differences between LLMs and human learners (see also Mitchell & Bowers, 2020). What differences there were could be explained by (non-linguistic) computational complexity of the "impossible" languages used. As Bowers (2025) puts it, "What makes a language difficult to learn is not the same for LLMs and humans, because they possess different inductive biases." More recently, Luo, Ramscar, and Love, (2024) and Ziv et al. (2025) found that LLMs did not distinguish between normal English and backward-reversed English text, again suggesting that, unlike the human language faculty (Moro, 2016) they make no distinction between possible and impossible languages.

The observation that LLMs learn to simulate impossible languages just as well as possible languages renders LLMs uninformative about the cognitive structures children use to acquire language. For one thing, children are incapable of acquiring linguistic patterns based on linear order, as Noam Chomsky has pointed out for decades (e.g., Chomsky, 1971). By contrast, the whole point about LLMs is that they can readily detect linear regularities in language input. This shows that LLMs and human children process language input in fundamentally different ways.

Given these fundamental differences between LLMs and the human language faculty, we fail to see how the former can inform the latter in any meaningful way. In summary, an LLM may quack like a duck, but isn't one – and, in its current probabilistic form, it can never be one.

Financial support. None.

Competing interests. None.

References

Bolhuis, J. J., Crain, S., & Roberts, I. (2023). Language and learning: The cognitive revolution at 60-odd. *Biological Reviews*, *98*, 931–941. https://doi.org/10.1111/brv.12936

Bolhuis, J. J., Crain, S., Fong, S., & Moro, A. (2024). AI doesn't model human language. *Nature*, *627*, 489. https://doi.org/10.1038/d41586-024-00824-z

Bowers, J. S. (2025). The successes and failures of artificial neural networks (ANNs) highlight the importance of innate linguistic priors for human language acquisition. *Psychological Review*, in press. Advance online publication. https://doi.org/10.1037/rev0000595

Chomsky, N. (1971). *Problems of knowledge and freedom: The Russell lectures*. Pantheon Books.

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

Chomsky, N. (2000). Minimalist inquiries: The framework. In R. Martin, D. Michaels, & J. Uriagereka (Eds.), *Step by step: Essays on minimalist syntax in honor of Howard Lasnik* (pp. 89–155). MIT Press.
Crain, S. (2012). *The emergence of meaning*. Cambridge University Press.
Crain, S., Koring, L., & Thornton, R. (2017). Language acquisition from a biolinguistic perspective. *Neuroscience & Biobehavioral Reviews*, *81*, 120–149. https://doi.org/10.1016/j.neubiorev.2016.09.004
Everaert, M. B., Huybregts, M. A., Chomsky, N., Berwick, R. C., & Bolhuis, J. J. (2015). Structures, not strings: Linguistics as part of the cognitive sciences. *Trends in Cognitive Sciences*, *19*, 729–743. https://doi.org/10.1016/j.tics.2015.09.008
Fong, S. (2025). The generative enterprise is alive. *Italian Journal of Linguistics*, *37*, 83–106. https://doi.org/10.26346/1120-2726-239
Friederici, A. D., Chomsky, N., Berwick, R. C., Moro, A. & Bolhuis, J. J. (2017). Language, mind and brain. *Nature Human Behaviour*, *1*, 713–722. https://doi.org/10.1038/s41562-017-0184-4
Kallini, J., Papadimitriou, I., Futrell, R., Mahowald, K., & Potts, C. (2024). Mission: Impossible language models. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd annual meeting of the association for computational linguistics (Volume 1: Long Papers)* (pp. 14691–14714). Association for Computational Linguistics. https://doi.org/10.48550/arXiv.2401.06416
Kerr, D. (2024). How Memphis became a battleground over Elon Musk's xAI supercomputer. NPR 11 September 2024. Retrieved from www.npr.org/2024/09/11/nx-s1-5088134/elon-musk-ai-xai-supercomputermemphis-pollution
Lawson, A. (2024). Google to buy nuclear power for AI datacentres in'world first' deal. Guardian 15 October 2024. Retrieved from www.theguardian.com/technology/2024/oct/15/google-buy-nuclear-power-ai-datacentreskairos-power
Ling, J. (2001). Power of a human brain. In *Physics factbook*. Retrieved from hypertextbook.com/facts/2001/JacquelineLing.shtml
Luo, X, Ramscar, M., & Love, B. C. (2024). Beyond human-like processing: Large language models perform equivalently on forward and backward scientific text. arXiv:2411.11061v1 [cs.CL]. https://doi.org/10.48550/arXiv.2411.11061
Markov, A. A. (1913/2006). An example of statistical investigation of the text Eugene Onegin concerning the connection of samples in chains. *Science in Context*, *19*(4), 591–600. https://doi.org/10.1017/S0269889706001074
Mitchell, J. & Bowers, J. (2020). Priorless recurrent networks learn curiously. In Proceedings of the 28th *International Conference* on *Computational Linguistics* (pp. 5147–5158). https://doi.org/10.18653/v1/2020.coling-main.451
Moro, A. (2016). *Impossible languages.* MIT Press.
Moro, A., Greco, M. & Cappa, S. F. (2023). Large languages, impossible languages and human brains. *Cortex*, *167*, 82–85. https://doi.org/10.1016/j.cortex.2023.07.003
Musso, M., Moro, A., Glauche, V., Rijntjes, M., Reichenbach, J., Büchel, C. & Weiller, C. (2003). Broca's area and the language instinct. *Nature Neuroscience*, *6*, 774–781. https://doi.org/10.1038/nn1077
Rizzi, L. (2009). The discovery of language invariance and variation, and its relevance for the cognitive sciences. *Behavioral and Brain Sciences*, *32*, 467–468. https://doi.org/10.1017/S0140525X09990574
Tettamanti, M., Alkadhi, H., Moro, A., Perani, D., Kollias, S., & Weniger, D. (2002). Neural correlates for the acquisition of natural language syntax, *NeuroImage*, *17*, 700–709. https://doi.org/10.1006/nimg.2002.1201
Yang, C., Crain, S., Berwick, R. C., Chomsky, N. & Bolhuis, J. J. (2017). The growth of language: Universal grammar, experience, and principles of computation. *Neuroscience & Biobehavioral Reviews*, *81*, 103–119. https://doi.org/10.1016/j.neubiorev.2016.12.023
Ziv, I., Chemmla, E., Lan, N., & Katzir, R. (2025). Biasless language models learn unnaturally: How LLM fail to distinguisch the possible from the impossible. arXiv:2510.07178v1 [cs.CL]. https://arxiv.org/abs/2510.07178v1 https://doi.org/10.48550/arXiv.2510.07178

Studies with impossible languages falsify LMs as models of human language

Jeffrey S. Bowers[a]* and Jeff Mitchell[b]

[a]School of Psychology and Neuroscience, University of Bristol, Bristol, UK and [b]Department of Informatics, University of Sussex, UK
j.bowers@bristol.ac.uk
j.j.mitchell@sussex.ac.uk

*Corresponding author.

doi:10.1017/S0140525X26104579, e202

Abstract

According to **F&M**, both infants and language models (LMs) find attested languages easier to learn than "impossible languages" that have unnatural structures. We review the literature and show that LMs often learn attested and many impossible languages equally well. Difficult to learn impossible languages are simply more complex (or random). LMs are missing human inductive biases that support language acquisition.

The speed with which infants learn languages is a core challenge for models of human language acquisition. Chomsky's hypothesis is that humans have an inductive bias, a "Universal Grammar" (UG), that not only constrains the linguistic structures of all attested languages, but also allows children to learn quickly. By contrast, language models (LMs), such as ChatGPT, lack human-like priors, and consequently, they display just the opposite profile, namely, they learn "impossible languages" that violate UG just as easily as attested ones (Mitchell & Bowers, 2020), and at the same time, require training on many orders of magnitude of more language data compared to children (Bowers, 2025a). According to Chomsky, *the ease with which LMs learn impossible languages undermines LMs as theories of human languages. For instance, Chomsky*,Roberts, and Watumul *(2023) write:*

> "Their deepest flaw is the absence of the most critical capacity of any intelligence: to say not only what is the case . . . but also what is not the case and what could and could not be the case . . .
>
> ChatGPT and similar programs are, by design, unlimited in what they can 'learn' (which is to say, memorize); they are incapable of distinguishing the possible from the impossible. Unlike humans, for example, who are endowed with a universal grammar that limits the languages we can learn to those with a certain kind of almost mathematical elegance, these programs learn humanly possible and humanly impossible languages with equal facility" [the underlined text encoded a link to Mitchell & Bowers, 2020].

Since Mitchell and Bowers (2020), there have been several studies that have compared how easily LMs learn attested and impossible languages. For instance, Kallini et al. (2024) reported two impossible languages that were more difficult to learn than English and concluded that LMs (and humans) do not need any linguistic priors to account for the possible-impossible learning gap. **F&M** describe this study and cite several others (including

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

Mitchell and Bowers) in a way that suggests that LMs consistently show a human-like difficulty in learning impossible languages, writing:

> "Kallini et al. (2024) find that the model learns from real English text consistently faster than these baselines (see also Mitchell & Bowers, 2020; Xu et al., 2025; Yang et al., 2025; Ziv et al., 2025, among others). These results and others show that Transformers and related models have inductive biases that align with human language."

But this characterization of the research is mistaken. First, consider the Kallini et al.'s study. Although the authors reported two impossible languages that were more difficult to learn, they reported that several additional impossible languages were learned almost as easily as English, including one of the impossible languages that Mitchell and Bowers (2020) devised. Furthermore, the impossible language that was the most difficult to learn was composed of random shuffles of words. That is, there was no structure to learn. The other difficult-to-learn impossible language was composed of deterministic random shuffles, with sentences of different lengths shuffled in different deterministic ways (e.g., all sequences of length 5 were shuffled in one order, sequences of length 6 were shuffled in another order, etc.). In this case, there was something to learn, namely, different random languages for the different sentence lengths. The observation that it is more difficult to learn multiple languages compared to a single random language does not challenge Chomsky's point. More generally, a learning algorithm that performs well on one data type must perform badly on others ("no free lunch theorems"; Wolpert & Macready, 2002). Accordingly, the observation that LMs perform badly on some languages is an unnecessary empirical confirmation of these theorems. In contrast, the success in learning some impossible languages is a falsification of the claim that LMs are an adequate model of human language learning.

The same issue applies to Yang et al. (2025) article that **F&M** cited. The authors again assessed LMs on a range of impossible languages, including the same deterministic shuffled languages used by Kallini et al. And again, the authors found that several impossible languages were easily learned, whereas the random shuffle language was difficult. And again, the slow learning of the random shuffled language was taken as problematic for Chomsky:

> "If LMs function as non-human like pattern recognizers as Chomsky et al. (2023); Moro et al. (2023) argue, they should be able to learn these languages as well as attested ones."

But again, the observation that it is more difficult to learn multiple different impossible languages (for different sentence lengths) is not surprising on any theory.

The Xu et al. (2025) article that **F&M** cited varied the plausibility of languages (rather than the impossibility of languages). In this case, the authors found that LMs struggled with some implausible languages, but again, in some cases, the LM found the implausible languages easy to learn. And with regard to conditions in which LMs struggled, the authors note that there may have been errors in the construction of the materials that weaken the conclusions that can be drawn:

> "Thus, it is possible that our findings may be due partially or entirely to increased noise in the counterfactual corpora, rather than inherent differences in learnability between the original and counterfactual grammars."

It is difficult to take these findings as problematic for Chomsky's thesis.

Finally, the Ziv et al. (2025) article that **F&M** cited reports several impossible languages that LMs learn just fine, including partially reversed languages (replicating Mitchell & Bowers, 2020). In addition, in another study not cited, Lou, Ramscar, and Love (2024) report that LMs can happily learn reversed languages.

So, contrary to the claims of **F&M**, LMs often learn impossible languages as easily as attested languages. At the same time, LMs require many orders of magnitude more training than infants (e.g., Bowers, 2025a), and a recent attempt to induce a universal grammar in LMs (McCoy & Griffiths, 2025) does not make LMs much more data efficient (Bowers, 2025b). Given all this, the appropriate conclusion is that current LMs learn in just the way one would expect of models that are missing human-like innate biases for language learning.

Acknowledgements. None.

Financial support. None.

Competing interests. None.

References

Bowers, J. S. (2025a). The successes and failures of artificial neural networks (ANNs) highlight the importance of innate linguistic priors for human language acquisition. Psychological Review. Advance online publication. https://doi.org/10.1037/rev0000595

Bowers, J. S. (2025b). Does distilling Bayesian priors into language models support rapid language learning? Comment on McCoy and Griffiths (2025). PsyArxiv. Preprint. https://doi.org/10.31234/osf.io/zeprf_v1

Chomsky, N. Roberts, I, & Watumull, J. (2023, March 8th). Noam Chomsky: The False Promise of ChatGPT. New York Times. https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html

Kallini, J., Papadimitriou, I., Futrell, R., Mahowald, K., & Potts, C. (2024). Mission: Impossible language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics* (Volume *1*: Long Papers, pp. 14691–14714). Association for Computational Linguistics

Luo, X., Ramscar, M., & Love, B. C. (2024). Beyond human-like processing: Large language models perform equivalently on forward and backward scientific text. arXiv preprint arXiv:2411.11061.

McCoy, R. T., & Griffiths, T. L. (2025). Modeling rapid language learning by distilling Bayesian priors into artificial neural networks. *Nature Communications*, *16*(1), 4676.

Mitchell, J., & Bowers, J. (2020, December). Priorless recurrent networks learn curiously. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 5147–5158).

Moro, A., Greco, M., & Cappa, S. F. (2023). Large languages, impossible languages and human brains. *Cortex*, *167*, 82–85.

Wolpert, D. H., & Macready, W. G. (2002). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, *1*(1), 67–82.

Xu, T., Kuribayashi, T., Oseki, Y., Cotterell, R., & Warstadt, A. (2025). Can language models learn typologically implausible languages? arXiv preprint arXiv:2502.12317.

Yang, X., Aoyama, T., Yao, Y., & Wilcox, E. (2025). Anything goes? A cross-linguistic study of (im) possible language learning in LMs. arXiv preprint arXiv:2502.18795

Ziv, I., Lan, N., Chemla, E., & Katzir, R. (2025). Large language models as proxies for theories of human linguistic cognition. arXiv preprint arXiv:2502.07687.

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

No usage-based linguistics without language use

Bastian Bunzeck[a]*, Sina Zarrieß[a], Stefan Hartmann[b] and Elena Lieven[c]

[a]Computational Linguistics, Linguistics Department, Bielefeld University, Bielefeld, Germany; [b]Department of German Languages and Literatures, Heinrich Heine University Düsseldorf, Düsseldorf, Germany and [c]Division of Psychology Communication and Human Neuroscience, The University of Manchester, Manchester, UK
bastian.bunzeck@uni-bielefeld.de
sina.zarriess@uni-bielefeld.de
hartmast@hhu.de
elena.lieven@manchester.ac.uk

*Corresponding author.

doi:10.1017/S0140525X26104646, e203

Abstract

Language models excel at finding patterns in linguistic data, and can therefore prove insightful for statistical approaches to linguistics in that they provide further evidence for the strong reliance of natural languages on recurrent, fixed patterns. Nevertheless, regarding actual usage-based language processing, their implications are severely limited as they lack a crucial aspect of language use: *interaction*.

Do language models model the language(s) of humans, and do they have anything to contribute to linguistic theory? This question has been debated by many (cf. Piantadosi, 2024 and its numerous replies), and Futrell and Mahowald offer one of the most optimistic answers: yes, language models do learn non-trivial linguistic structure, and linguists should therefore recognize their value. An essential part of their argument draws on statistical and usage-based traditions in linguistics, which highlight that human language learning relies heavily on distributional patterns. On this view, grammar in the human mind is best understood not as a system of purely symbolic rules, but as a repertoire of structures shaped by probabilities and frequencies. Put simply, Futrell and Mahowald take modern LMs as vindicating these "statistical approaches" to language, since such models learn language strikingly well from raw text alone, without being endowed with strong inductive biases or constraints. However, in emphasizing how language models serve as a proof of concept for many of their core claims, Futrell and Mahowald paint a reductionist picture of usage-based approaches. It is certainly true that current (L)LMs are compatible with usage-based ideas, and we concur with their idea that LMs may point to "real patterns" in the sense of Dennett (1989) that can also prove insightful for usage-based linguistics. Yet, at the same time, they entirely lack the socio-cognitive grounding that usage-based linguistics considers constitutive of language. Therefore, the implications of current language models for usage-based linguistics are severely limited, despite widespread claims to the contrary (e.g., Ambridge & Blything, 2024, Goldberg, 2024, Piantadosi, 2024).

Like human learners, LMs excel at finding patterns in data and learn to produce fluent language through domain-general statistical mechanisms. But Futrell and Mahowald's focus on frequencies and distributions sidelines much of what usage-based linguists are actually interested in: the ways in which cognitive, social, and interactive pressures shape the linguistic system and how these factors enable humans to use language as fluently as they do. Language models learn and process language input unlike humans. They learn from statistical associations in corpora, whereas humans learn from associations of linguistic forms and their meanings in social contexts and their lived experience. From a usage-based point of view, learning means "constructing new representations via interaction with environmental stimuli" (Rowland et al., 2025). This learning process rests on a number of domain-general cognitive capacities, including the registration of frequencies and transitional probabilities. But crucially, it also relies on social-interactional environmental contexts (Enfield, 2017) to provide the mapping of meaning to form. Intention reading and, in many contexts, joint attention are crucial to identifying the meaning of linguistic forms and to learning their mappings (Ibbotson, 2011; Tomasello, 2003).

Hence, for human language, the relevance of interaction, grounding in the real world, and the functional dimension of language can hardly be underestimated. Putting these processes into the center of linguistic inquiry is a core tenet of usage-based linguistics, and certain statistical patterns might only emerge as side effects of these processes. However, they are not represented in the lifecycle of current, widely used language models, which are typically trained on gigantic corpora, much larger than the linguistic input that humans typically receive over their lifespan (cf. Brysbaert et al., 2016). Initial pre-training on unstructured data is complemented by mid-training on reasoning data as well as post-training/preference tuning on task-oriented dialogs. Moreover, current evaluation methods probing the linguistic capabilities of language models remain anchored in formalizations of language perception. For example, the BabyLM test suite (Charpentier et al., 2025) contains, *inter alia*, minimal pair tests where preference for supposedly "grammatical" sentences is measured through string probabilities (BLiMP, Warstadt et al., 2020), and correlation tests for word-level surprisal and human acquisition norms (Chang & Bergen, 2022). These tests correspond to deliberate experimental interventions in humans, not their natural language use. Recent benchmarks embrace fewer binary assumptions of language structure, for example, by comparing language models' preferences for different nonce word inflections with instances of human usage (Hofmann et al., 2025). Still, these tests detach human language from its socio-cognitive context. Evaluation methods that test for functional aspects of language, like social and real-world knowledge (e.g., SiQA, Sap et al., 2019; EWoK, Ivanova et al., 2025), necessarily cast them as text-based problems, on which language models regularly underperform (Mahowald et al., 2024).

In fact, even the very notion of language (use) is fundamentally different between language models and humans. For a language model, language is a probability distribution over sets of strings. In usage-based linguistics, using language is a way of acting in the social world, something that language models cannot do. A nice example is provided by Engelmann et al. (2019), who compared the learning of person-number agreement by Finnish and Polish children with that of a computational model. While the order of acquisition was similar for both, there was one exception. Despite the very high frequency of the second-person singular in the input, children produced the self-referring first-person singular earlier and more accurately, while, by contrast, the model simply reflected the frequencies in the input and produced the second-person singular well before the first-person singular.

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

In light of these crucial differences between language (use) in humans and language models, we argue that (L)LMs remain poor objects of linguistic inquiry from a usage-based perspective. As of now, they cannot model "the real thing" (p. 13), simply because the real thing is not just a (highly apt) text-generation mechanism, but a complex, embodied, context-sensitive process shaped by interaction and social experiences. This problem has been partially recognized by the functionally inspired language modeling community. There, reinforcement learning is being applied with more developmentally plausible objectives, such as communicative success in interaction (e.g., Mayer Martins et al., 2025; Padovani et al., 2025; Salhan et al., 2025). Still, most of these experiments show that optimization for such goals leads to deteriorating scores on linguistic benchmarks and impairs text generation capabilities. The aim of modeling more realistic language use seems to be broadly at odds with current LM architectures, and it remains unclear if (and how) these aims could be reconciled.

Acknowledgements. None.

Financial support. BB and SZ are supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – CRC-1646, project number 512393437, project A02. SH is supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), project number 504095269.

Competing interests. None.

References

Ambridge, B., & Blything, L. (2024). Large language models are better than theoretical linguists at theoretical linguistics. *Theoretical Linguistics*, *50*(1–2), 33–48. https://doi.org/10.1515/tl-2024-2002

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. *Frontiers in Psychology*, *7*, 1–22. https://doi.org/10.3389/fpsyg.2016.01116

Chang, T. A., & Bergen, B. K. (2022). Word acquisition in neural language models. *Transactions of the Association for Computational Linguistics*, *10*, 1–16. https://doi.org/10.1162/tacl_a_00444

Charpentier, L., Choshen, L., Cotterell, R., Gul, M. O., Hu, M. Y., Liu, J., Jumelet, J., Linzen, T., Mueller, A., Ross, C., Shah, R. S., Warstadt, A., Wilcox, E. G., & Williams, A. (2025). Findings of the third BabyLM challenge: Accelerating language modeling research with cognitively plausible data. Proceedings of the First BabyLM Workshop, 399–420. https://doi.org/10.18653/v1/2025.babylm-main.28

Dennett, D. C. (1989). *The intentional stance*. MIT Press.

Enfield, N. J. (2017). Opening commentary: Language in cognition and culture. In B. Dancygier (Ed.), *The Cambridge handbook of cognitive linguistics* (pp. 13–18). Cambridge University Press. https://doi.org/10.1017/9781316339732.002

Engelmann, F., Granlund, S., Kolak, J., Szreder, M., Ambridge, B., Pine, J., Theakston, A., & Lieven, E. (2019). How the input shapes the acquisition of verb morphology: Elicited production and computational modelling in two highly inflected languages. *Cognitive psychology*, *110*, 30–69. https://doi.org/10.1016/j.cogpsych.2019.02.001

Goldberg, A. E. (2024). Usage-based constructionist approaches and large language models. *Constructions and Frames*, *16*(2), 220–254. https://doi.org/10.1075/cf.23017.gol

Hofmann, V., Weissweiler, L., Mortensen, D. R., Schütze, H., & Pierrehumbert, J. B. (2025). Derivational morphology reveals analogical generalization in large language models. *Proceedings of the National Academy of Sciences*, *122*(19), e2423232122. https://doi.org/10.1073/pnas.2423232122

Ibbotson, P. (2011). Abstracting grammar from social–cognitive foundations: A developmental sketch of learning. *Review of General Psychology*, *15*, 331–343. https://doi.org/10.1037/a0025609

Ivanova, A. A., Sathe, A., Lipkin, B., Kumar, U. U., Radkani, S., Clark, T. H., Kauf, C., Hu, J., Pramod, R. T., Grand, G., Paulun, V. C., Ryskina, M., Akyürek, E., Wilcox, E. G., Rashid, N., Choshen, L., Levy, R., Fedorenko, E., Tenenbaum, J., & Andreas, J. (2025). Elements of world knowledge (EWOK): A cognition-inspired framework for evaluating basic world knowledge in language models. *Transactions of the Association for Computational Linguistics*, *13*, 1245–1270. https://doi.org/10.1162/TACL.a.38

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2024). Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, *28*(6), 517–540. https://doi.org/10.1016/j.tics.2024.01.011

Mayer Martins, J., Bashir, A. H., Khalid, M. R., & Beinborn, L. (2025). Once upon a time: Interactive learning for storytelling with small language models. Proceedings of the First BabyLM Workshop, 454–468. https://doi.org/10.18653/v1/2025.babylm-main.32

Padovani, F., Bunzeck, B., Ali, M., Momen, O., Bisazza, A., Buschmeier, H., & Zarrieß, S. (2025). Dialogue is not enough to make a communicative BabyLM (But neither is developmentally inspired reinforcement learning). Proceedings of the First BabyLM Workshop, 421–435. https://doi.org/10.18653/v1/2025.babylm-main.29

Piantadosi, S. T. (2024). Modern language models refute Chomsky's approach to language. In E. Gibson, & M. Poliak (Eds.), *From fieldwork to linguistic theory. A tribute to Dan Everett* (No. 15; pp. 353–414). Language Science Press. https://doi.org/10.5281/zenodo.11351540

Rowland, C. F., Westermann, G., Theakston, A. L., Pine, J. M., Monaghan, P., & Lieven, E. V. M. (2025). Constructing language: A framework for explaining acquisition. Trends in Cognitive Sciences, S1364661325001421. https://doi.org/10.1016/j.tics.2025.05.015

Salhan, S., Gu, H., Rooein, D., Galvan-Sosa, D., Gaudeau, G., Caines, A., Yuan, Z., & Buttery, P. (2025). Teacher demonstrations in a BabyLM's zone of proximal development for contingent multi-turn interaction. Proceedings of the First BabyLM Workshop, 323–355. https://doi.org/10.18653/v1/2025.babylm-main.25

Sap, M., Rashkin, H., Chen, D., Le Bras, R., & Choi, Y. (2019). Social IQA: Commonsense reasoning about social interactions. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 4462–4472. https://doi.org/10.18653/v1/D19-1454

Tomasello, M. (2003) *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., & Bowman, S. R. (2020). Blimp: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, *8*, 377–392. https://doi.org/10.1162/tacl_a_00321

On the relationship between theoretical and empirical research in linguistics

Luca Capone and Alessandro Lenci*

CoLing Lab, Department of Philology, Literature and Linguistics, University of Pisa, Italy
luca.capone@fileli.unipi.it
alessandro.lenci@unipi.it

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

*Corresponding author.

doi:10.1017/S0140525X26104774, e204

Abstract

Language models (LMs) call for a theoretical rethinking grounded within linguistics itself. Rather than signalling the "end of linguistics" or merely encouraging interdisciplinarity, LMs function as an empirical testing ground for formal linguistic concepts. They prompt a renewed examination of form–meaning mapping, theoretical autonomy, and the conditions under which computational systems can genuinely inform linguistic explanation.

The supposed "end of linguistic theory," allegedly superseded by language models (LMs), has been a recurring topic of debate. Futrell and Mahowald (F&M) offer a different perspective, one that is more conciliatory and readily acceptable: LMs do not constitute a complete theory of language, but they provide a wealth of information about it. Given this, what should be the scope and aims of linguistic theory in the era of "triumphing LMs," and how should the two domains be properly related?

F&M point out that LMs are "models of the stream of language produced and comprehended by humans." However, strictly speaking, they are even less than that: they are just models of the stream of *linguistic forms* produced by humans, to which they attach semantic content to the extent this is recoverable from distributional statistics. After all, as they correctly say, LMs are the most advanced product of distributional semantics (Lenci & Sahlgren, 2023). On the other hand, the goal of linguistics should be more than *form sequence modelling*. Its aim has always been to model natural language as a potentially infinite system of *signs* intended as *form-meaning* pairs (de Saussure, 1916). Linguistic theory must therefore be a theory of how meaning representations are mapped back and forth onto linguistic forms for the purpose of communication. Similarly, the goal of a theory of language learning is to model how this mapping is acquired from a finite and relatively small amount of communicative experiences. Generative linguistics might be limited in pursuing this goal only with reference to language as an abstract system, as F&M argue, but other linguistic theories that take a very different perspective, such as the Parallel Architecture (Jackendoff, 2002) and Construction Grammar (Goldberg, 2019), also assume that the aim of linguistics is modelling language as form-meaning mappings. Therefore, however substantial the potential contribution of LMs may be, it must be understood within this broader goal of linguistic theory.

F&M advocates a "decentralised" approach to linguistic inquiry, according to which partial insights drawn from other sciences and technologies, such as LMs, will gradually reveal the contours of the linguistic object, like an unknown continent to be charted or a jigsaw puzzle to be assembled. As they write, "We will not answer the most fundamental questions about language by focusing *only* on language, using *only* methods that are tailor-made for linguistics." Within this perspective, the role of linguistic theory might be intended as just one of *reverse engineering*: The inner working of the first "language machines" will lead to understanding how language works. Although we concur with the claim that some insights about language will come from the analysis of LMs, we believe this is not the best possible setting for a fruitful dialogue between linguistics and LMs, given the very nature of linguistics and its goals, as stated above.

First of all, linguistics occupies a special status among the sciences. According to De Saussure, what distinguishes linguistics is precisely that its theoretical viewpoint constitutes its own object of inquiry. This principle is also shared by Zellig Harris (1991), one of the founders of the distributional paradigm that grounds the way in which LMs acquire linguistic categories. Unlike physics or biology, linguistics does not confront pre-existing natural objects, it creates its own epistemic domain through concepts such as *sign*, *structure*, and *meaning*. The search for physiological or psychological correlates of linguistic phenomena in the brain and mind already presupposes a prior definition of the linguistic entity for which the correlate is sought. Whatever physiological or psychological phenomenon may be observed, it cannot, by definition, constitute a linguistic phenomenon, which is linguistic in nature by the very premises of the inquiry. Similarly, the reverse engineering of LMs requires the analysis of how linguistic categories are encoded in their architectures, which in turn presupposes a prior definition of the linguistic entities and structures we aim to seek in the models' internal states.

Secondly, deriving a scientific theory about a certain function from a machine's behaviour is only possible if this function has been truly realised by the machine. The physical theory of aerodynamics has greatly benefited from the study of airplanes, but this was possible because such machines were indeed able to fly. As we said above, LMs are just machines that produce strings of linguistic forms that closely approximate the ones produced by humans, but do not really master language in the true sense of performing a mapping between forms and meanings. For instance, as Mahowad et al. (2024) convincingly argue, there is a mismatch between the formal and the functional competences of LMs. Actually, this mismatch is not a secondary issue, but a necessary consequence of there being distributional devices that piggyback all meaning onto form co-occurrences. We might regard it as another case of the "embers of autoregression" of LMs *qua* LMs (McCoy et al., 2023).

What then is the role of LMs for linguistic theory? We agree with F&M that LMs have great potential for linguistics. We think that LMs provide a new "tribunal of experience" (Quine, 1953) to which linguistic theory must be submitted. However, it does not follow from this that we should develop linguistic theories just by reverse engineering LMs. It is instead necessary that independently developed hypotheses and theories about language be validated on computational models like LMs. In this light, LMs can become a powerful means for falsifying core conceptual systems that make linguistics a science in the first place. However, this is possible only under two conditions: 1.) linguistic theory keeps its autonomy from LMs, very much like theoretical physics is independent from the experimental methods validating it; 2.) linguistics and LMs must grow closer than they currently are. This calls for a reciprocal effort: Linguistics must accept LMs as empirical judges of its theories, but LMs must increase their plausibility as experimental testbeds for natural language models in the way they are trained and designed. If this encounter happens, the dialogue will be incredibly beautiful.

Acknowledgments. None.

Financial support. We acknowledge financial support under the PRIN 2022 Project Title "Computational and linguistic benchmarks for the study of verb argument structure" – CUP I53D23004050006 – Grant Assignment Decree No. 1016 adopted on 07/07/2023 by the Italian Ministry of University and Research (MUR). This work was also supported under the PNRR – M4C2 – Investimento 1.3, Partenariato Esteso PE00000013 – FAIR – Future Artificial Intelligence

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

Research – Spoke 1 "Human-centered AI," funded by the European Commission under the NextGeneration EU programme.

Competing interests. None.

References

Goldberg, A. E. (2019). *Explain me this. Creativity, competition, and the partial productivity of constructions*. Princeton University Press.

Harris, Z. S. (1991). *A theory of language and information: A mathematical approach*. Clarendon Press.

Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar, evolution*. Cambridge University Press.

Lenci, A., & Sahlgren, M. (2023). *Distributional semantics*. Cambridge University Press.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2024). Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, *28*(6), 517–540.

McCoy, R. T., Yao, S., Friedman, D., Hardy, M., & Griffiths, T. L. (2023). Embers of autoregression: Understanding large language models through the problem they are trained to solve. *ArXiv*:2309.13638.

Quine, W. V. O. (1980 [1953]). Two Dogmas of empiricism. In W.V.O Quine (Eds.), *From a logical point of view* (2nd ed., pp. 20–46). Harvard University Press.

de Saussure, F. (2011 [1916]). *Course in general linguistics*. Columbia University Press.

Language models do not yet achieve explanatory adequacy because language is more than incremental prediction

Jennifer Culbertson*, Simon Kirby and Kenny Smith

The University of Edinburgh, UK
jennifer.culbertson@ed.ac.uk
simon.kirby@ed.ac.uk
kenny.smith@ed.ac.uk

*Corresponding author.

doi:10.1017/S0140525X26104683, e205

Abstract

In this commentary, we challenge the idea that Language Models (LMs) provide explanatorily adequate models of human language. Findings from language evolution show that both humans and LMs fail to produce human-like language via inductive biases alone; the communicative function of language is crucial. More generally, both experimental and computational work underscore the fact that language is more than just incremental prediction.

As linguists working on language evolution, we share F&M's enthusiasm about the potential for Language Models (LMs) to help answer questions about language learning and the human language faculty. Studying the ability of artificial systems like LMs to acquire important features of natural languages is a potentially powerful tool for identifying critical cognitive capacities that contribute to language learning in humans. We see parallels with the long history of work in comparative psychology, uncovering capacities in humans that may be shared with non-human animals (see Arnon et al., 2025; Gibson, 2011; Petkov & ten Cate, 2020). However, current LMs far outstrip any non-human animal in their ability to approximate human language. As F&M point out, to the extent that LMs assign structures to sentences that align with human intuition and generalise in the same way that humans do, they provide an existence proof that linguistic input alone is sufficient to achieve descriptive adequacy for human language.

But as F&M also point out, a central goal of linguistics is *explanatory* adequacy, i.e., explaining how language is learned and why some logically possible types of language are rare or unattested. F&M, following Chomsky, assume that learnability and language universals are straightforwardly causally linked: constraints on language variation arise from inductive biases, and inductive biases active during learning will also lead LMs to generate only human-like languages. We would like to challenge this assumption. Learning alone cannot explain language structure, because learning is only one of the constraints acting on the cultural process by which languages are transmitted.

F&M argue that both humans and LMs "*encode and decode information in linguistic strings incrementally and predictively . . . To the extent that human language is structured in a way that supports this kind of incremental prediction . . . it is also aligned with the . . . inductive biases of language models.*" In other words, shared functions can lead to convergent, shared solutions. F&M makes the analogy between bird and airplane wings: both are shaped by the function of flight. However, this is at best only one highly proximal function that shapes these systems. A theory of bird flight based on modern airplane design would fail to account for the many differences between these two methods of flying through the air. To account for the differences between birds and planes, we need to consider less proximate functions than just aerodynamic lift, such as survival and reproduction in the case of the bird, and transporting passengers at low cost in the case of the plane. Similarly, LMs and human language are shaped by the function of incremental prediction, but this captures only one narrow aspect of communicative function. After all, human language is not purely a product of learning to incrementally predict. It is also shaped by the need to activate and transform conceptual representations, including complex thoughts, grounded in world knowledge and reflecting features of human perception and action, into language so that they can be shared with others (e.g., Culbertson et al., 2020). In other words, LMs are trained only to incrementally predict language, whereas for humans, that task is incidental, or only necessary so far as it facilitates communication. Any theory of language that does not explicitly incorporate the communicative function of language will thus not achieve explanatory adequacy; it will fail to account for constraints on human language arising from this function.

To the extent that LMs, tasked with next word prediction, produce human-like language, they may do this because they are trained on data generated by humans with real communicative goals, not because they themselves have communicative goals. In the absence of a communicative task, even human learners will produce language that is highly predictable but not communicatively functional. To explore this, we use laboratory experiments in which human participants learn and reproduce a miniature language, thereby transmitting it over multiple simulated generations. If these languages are not used for communication, then the languages that evolve in this cultural process are shaped purely by pressures for learnability, ultimately leading to simple (i.e., easy-to-

learn) systems which are useless for communication (e.g. Kirby, Cornish, & Smith, 2008). Having participants also use the miniature languages for communication provides a crucial counteracting pressure, preventing this kind of language collapse. Only when both learning and communication are part of the cultural process do we see structured systems emerge with language-like properties (e.g., Kirby et al., 2015; Motamedi et al., 2019; Smith & Culbertson, 2025). Similarly, if LMs are trained on text generated by other LMs, languages transmitted in this way degenerate as they are passed from model to model; later models produce probable sentences too frequently and generate meaningless sequences that no human would ever produce (Shumailov et al., 2024). F&M do mention another body of promising modelling work, the so-called emergent communication paradigm (e.g., Chaabouni et al., 2021; Havrylov & Titov, 2017; Lazaridou, Peysakhovich, & Baroni, 2017). This does explicitly model communication, usually between pairs of simulated agents, but uses different model architectures and learning algorithms, not based on incremental prediction, which do not currently come close to achieving descriptive adequacy for human language.

Both iterated learning experiments and artificial language learning experiments, in which participants must communicate and extrapolate beyond their input, have been particularly helpful in building explanatory theories of human language (e.g., Culbertson, 2023; Smith, 2022). To the extent that languages emerging from similar experiments with LMs differ from the features we see in human populations, then the particularities of human cognition, human inductive biases, human communicative goals, and/or human non-linguistic experience actually matter for explaining language structure. This is an exciting open question, and we look forward to more research that uses methods from language evolution to probe the explanatory adequacy of LMs. Particularly promising for us is to manipulate the input and training objectives to incorporate more of what we have suggested might currently be missing: world/conceptual knowledge and representations, and the task of mapping between these representations and language, in the context of communication.

Financial support. This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Competing interests. None.

References

Arnon, I., Carmel, L., Claidière, N., Fitch, W. T., Goldin-Meadow, S., Kirby, S., Okanoya, K., Raviv, L., Wolters, L., & Fisher, S. E. (2025). What enables human language? A biocultural framework. *Science*, *390*(6775), eadq8303.

Chaabouni, R., Kharitonov, E., Dupoux, E., & Baroni, M. (2021). Communicating artificial neural networks develop efficient color-naming systems. *Proceedings of the National Academy of Sciences*, *118*(12), e2016569118.

Culbertson, J. (2023). Artificial language learning. *Oxford Handbook of Experimental Syntax*, Chapter 8, 271–300.

Culbertson, J., Schouwstra, M., & Kirby, S. (2020). From the world to word order: Deriving biases in noun phrase order from statistical properties of the world. *Language*, *96*(3), 696–717.

Gibson, K. R. (2011). Language or protolanguage? A review of the ape language literature. In K. R. Gibson, & M. Tallerman (Eds.), *The oxford handbook of language evolution* (pp. 46–58). Oxford University Press.

Havrylov, S., & Titov, I. (2017). Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 2146–2156.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*, 10681–10686.

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87–102.

Lazaridou, A., Peysakhovich, A., & Baroni, M. (2017). Multi-agent cooperation and the emergence of (natural) language. In *5th International Conference on Learning Representations, ICLR 2017*, Toulon, France.

Motamedi, Y., Schouwstra, M., Smith, K., Culbertson, J., & Kirby, S. (2019). Evolving artificial sign languages in the lab: From improvised gesture to systematic sign. *Cognition*, *192*, 103964.

Petkov, C. I., & ten Cate, C. (2020). Structured sequence learning: Animal abilities, cognitive operations, and language evolution. *Topics in Cognitive Science*, *12*, 828–842

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024) AI models collapse when trained on recursively generated data. *Nature*, *631*, 755–759.

Smith, K. (2022). How language learning and language use create linguistic structure. *Current Directions in Psychological Science*, *31*, 177–186.

Smith, K., & Culbertson, J. (2025). Communicative pressures shape language during communication (not learning): Evidence from case-marking in artificial languages. *Cognition*, *263*, 106164.

Linguists should learn to love speech-based deep learning models

Marianne de Heer Kloots[a]*, Paul Boersma[b] and Willem Zuidema[a]

[a]Institute for Logic, Language and Computation (ILLC), University of Amsterdam, Amsterdam, Netherlands and [b]Amsterdam Center for Language and Communication (ACLC), University of Amsterdam, Amsterdam, Netherlands
m.l.s.deheerkloots@uva.nl
paul.boersma@uva.nl
w.h.zuidema@uva.nl

*Corresponding author.

doi:10.1017/S0140525X26104713, e206

Abstract

Futrell and Mahowald present a useful framework bridging technology-oriented deep learning systems and explanation-oriented linguistic theories. Unfortunately, the target article's focus on generative text-based Large Language Models (LLMs) fundamentally limits fruitful interactions with linguistics, as many interesting questions on human language fall outside what is captured by written text. We argue that audio-based deep learning models can and should play a crucial role.

To integratively study a human spoken language, linguists investigate how speakers map their communicative intent all the way to their articulatory gestures (language production), as well as how listeners map incoming auditory signals to their interpretation of a speaker's intent (language comprehension). An inclusive linguistics is therefore majorly concerned with structural relations

between physical signals and linguistic content – relations typically studied by the fields of *phonetics* (audition, acoustics, and articulation) and *phonology* (sound structure). As these fields are largely irrelevant to text-based Large Language Models (LLMs), F&M don't address them. This jeopardizes their own endeavour.

The trouble with text

For language comprehension, LLMs operate on text that has already discretized the continuous auditory speech stream into things that look like words, morphemes, syllables, and/or phonological segments (vowels and consonants), depending on the language. Conversely, crucial phonological aspects of spoken language like prosody and intonation aren't usually represented in text. As a result, ambiguities differ between text and speech: English text but not speech distinguishes between *sun* and *son*, while English speech but not text disambiguates the polar vs. alternative readings of *Do you like coffee or tea?* In general, text-based models end up solving fundamentally different problems than human spoken language users, as both cognitive scientists and language technologists have noted before (Chrupała, 2023; Dupoux, 2018).

The structures of natural languages reflect overwhelmingly the properties of speech (or signing), rather than those of text. Merely enriching text-based LLMs with speech capacities through acoustic tokens or separate speech recognition/synthesis components (Arora et al., 2025) cannot resolve the deep trouble caused by textual bottlenecks in linguistic modelling. Researchers interested in modelling linguistic structure should therefore not settle for LLMs but rather learn to love models designed to capture the more natural form of human spoken language: the speech signal itself.

Linguistic structure in models of speech

Many current speech-technological applications employ self-supervised *speech foundation models*, i.e., deep learning architectures trained to represent audio signals on the basis of unlabelled speech recordings. Parallel to work on the interpretability of text-based LLMs (as covered by F&M), *linguistic interpretability* studies in the speech domain investigate to what extent speech-based models learn to capture any higher-level patterns that make up spoken language. Method-wise, most of these studies use *representational probes* to examine what linguistic information is represented in speech models' internal states, with insightful results obtained for the encoding of linguistic units like phonemes (Alishahi, Barking, & Chrupała, 2017; Martin et al., 2023) and words (Pasad et al., 2024), as well as morphophonological (Gauthier et al., 2025) and suprasegmental patterns (Shen et al., 2024); other studies use *behavioural (minimal pair) tests* to ask, e.g., whether models can distinguish words from pseudowords (Lavechin et al., 2023) or prosodically natural vs. unnatural pauses (de Seyssel et al., 2023). With these types of methods, speech models can be used as "psycholinguistic participants," mimicking human speech perception experiments and their results, e.g., about perceptual similarity (Millet & Dunbar, 2022) and phonetic categorization (de Heer Kloots & Zuidema, 2024).

Linguists can already draw an important theoretical insight from these findings: speech-only models can indeed learn relevant patterns of linguistic structure, without the pre-existing symbolic categories assumed by earlier connectionist work (McClelland & Elman, 1986; Smolensky, 1990, 1999).

Inductive biases from domain-general perception

Further investigations of what deep learning models learn from audio can potentially inform what mechanisms drive language-relevant perceptual learning in humans. Here, we agree with F&M that deep learning models can helpfully serve as proof-of-concepts for ideas that were previously hard to formalize. At least part of the human perceptual system involved in speech processing is also involved in processing a wide variety of other auditory signals. Deep learning models pre-trained on music and/or environmental sounds show more human-like behaviour than models trained on speech sounds alone, in detecting algebraic patterns in tone or syllable sequences (Orhan, Boubenec, & King, 2025) and in displaying native language effects when processing foreign speech sounds (Poli et al., 2024). Hence, it seems that some inductive biases relevant to the encoding of language-like structures can result from (evolutionarily) optimizing the perceptual system for representing sounds other than speech, perhaps more so than considered before (Aslin & Pisoni, 1980; Soderstrom, Mathis, & Smolensky, 2006).

Inductive biases from bidirectional processing

One property of much cognitively grounded neural modelling, that is not shared with current technological speech or text machines, is that connection weights are symmetric (Kohonen, 1982; Hopfield, 1982; McMurray et al., 2009; Salakhutdinov & Hinton, 2009). In linguistics, this corresponds to a specific inductive bias, namely that language employs *bidirectional processing*, i.e., language comprehension and language production utilize the same knowledge.

In phonetics and phonology, neural models with bidirectional connection weights straightforwardly predict the diachronic evolution of auditory dispersion in phoneme inventories (Boersma, Benders, & Seinhorst, 2020). Unpublished simulations show that such communicative-success-optimizing effects readily transfer to other linguistic subdomains: with bidirectional connection weights, the maxims of Grice (1975) emerge in pragmatics, as do anti-synonymity effects in semantics, morphology, and syntax (Boersma, 2009).

A drawback of small neural models is that they work exclusively on *toy problems*, rather than *whole languages*. Hopefully, evidence from toy models can nevertheless inspire speech technologists in architectural design. Most findings cited above were obtained with speech models trained on developmentally realistic amounts of speech (< 1000 hours); beyond improving data efficiency, exploring a variety of inductive biases is primarily crucial in developing more human-like systems with relevance for linguistic theory.

Take-home message

We agree that neural network models developed for technological applications can provide insights into human language and cognition. The early connectionist work that F&M describe as preceding the LLM era included efforts to take the continuous speech signal seriously (Elman & Zipser, 1988; Norris, 1994), and we argue that current linguistic investigations with deep learning models can and should do the same. Once the bottleneck of text can be replaced, we stand a better chance of modelling more human-like linguistic processes, as well as handling the vast majority of local

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

and regional language varieties that are rarely or never written down.

Financial support. MdHK is funded by the Netherlands Organization for Scientific Research (NWO), through a Gravitation Grant 024.001.006 to the Language in Interaction Consortium. Open access funding provided by University of Amsterdam.

Competing interests. The authors declare none.

References

Alishahi, A., Barking, M., & Chrupała, G. (2017). Encoding of phonology in a recurrent neural model of grounded speech. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)* (pp. 368–378). https://doi.org/10.18653/v1/K17-1037

Arora, S., Chang, K. W., Chien, C. M., Peng, Y., Wu, H., Adi, Y., Dupoux, E., Lee, H. Y., Livescu, K., & Watanabe, S. (2025). On the landscape of spoken language models: A comprehensive survey. *arXiv preprint*. https://doi.org/10.48550/arXiv.2504.08528

Aslin, R., & Pisoni, D. (1980). Some developmental processes in speech perception. In G. H. Yeni-Komshian, J. F. Kavanagh, & C. A. Ferguson (Eds.), *Child phonology: Vol 2, Perception* (pp. 67–96). Academic Press.

Boersma, P. (2009). Unidirectional optimization of comprehension can achieve bidirectional optimality. Talk presented at 10th Szklarska Poręba Workshop on the Roots of Pragmasemantics, Szklarska Poręba, March 13, 2009.

Boersma, P., Benders, T. & Seinhorst, K. (2020). Neural networks for phonology and phonetics. *Journal of Language Modelling*, *8*(1), 103–177. https://doi.org/10.15398/jlm.v8i1.224

Chrupała, G. (2023). Putting Natural in Natural Language Processing. In *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 7820–7827). https://doi.org/10.18653/v1/2023.findings-acl.495

de Heer Kloots, M., & Zuidema, W. (2024) Human-like linguistic biases in neural speech models: Phonetic categorization and phonotactic constraints in Wav2Vec2.0. Proc. Interspeech 2024, 4593–4597. https://doi.org/10.21437/Interspeech.2024-2490

de Seyssel, M., Lavechin, M., Titeux, H., Thomas, A., Virlet, G., Revilla, A.S., Wisniewski, G., Ludusan, B., Dupoux, E. (2023) ProsAudit, a prosodic benchmark for self-supervised speech models. Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH 2023), 2963–2967. https://doi.org/10.21437/Interspeech.2023-438

Dupoux, E. (2018). Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, *173*, 43–59. https://doi.org/10.1016/j.cognition.2017.11.008

Elman, J. L., & Zipser, D. (1988). Learning the hidden structure of speech. *The Journal of the Acoustical Society of America*, *83*(4), 1615–1626. https://doi.org/10.1121/1.395916

Gauthier, J., Breiss, C., Leonard, M. K., & Chang, E. F. (2025). Emergent morpho-phonological representations in self-supervised speech models. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (pp. 28055–28074). https://doi.org/10.18653/v1/2025.emnlp-main.1425

Grice, H. P. (1975). Logic and conversation. In P. Cole, & J.J. Morgan (Eds.), *Syntax and semantics 3: Speech acts* (pp. 41–58). Academic Press.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, *79*, 2554–2558. https://doi.org/10.1073/pnas.79.8.2554

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, *43*, 59–69. https://doi.org/10.1007/BF00337288

Lavechin, M., Sy, Y., Titeux, H., Blandón, M.A.C., Räsänen, O., Bredin, H., Dupoux, E., Cristia, A. (2023). BabySLM: Language-acquisition-friendly benchmark of self-supervised spoken language models. Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH 2023), 4588–4592. http://doi.org/10.21437/Interspeech.2023-978

Martin, K., Gauthier, J., Breiss, C., Levy, R. (2023) Probing self-supervised speech models for phonetic and phonemic information: A case study in aspiration. Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH 2023), 251–255. http://doi.org/10.21437/Interspeech.2023-2359

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*(1), 1–86. https://doi.org/10.1016/0010-0285(86)90015-0

Millet, J., & Dunbar, E. (2022). Do self-supervised speech models develop human-like perception biases?. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 7591–7605). https://doi.org/10.18653/v1/2022.acl-long.523

McMurray, B., Horst, J. S., Toscano, J. C., & Samuelson, L. K. (2009). Integrating connectionist learning and dynamical systems processing: Case studies in speech and lexical development. In J. P. Spencer, M. S. C. Thomas, & J. L. McClelland (Eds.), *Toward a unified theory of development: Connectionism and dynamic systems theory re-considered* (pp. 218–249). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195300598.003.0011

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition*, *52*(3), 189–234. https://doi.org/10.1016/0010-0277(94)90043-4

Orhan, P., Boubenec, Y., & King, J. R. (2025). The detection of algebraic auditory structures emerges with self-supervised learning. *PLOS Computational Biology*, *21*(9), e1013271. https://doi.org/10.1371/journal.pcbi.1013271

Pasad, A., Chien, C. M., Settle, S., & Livescu, K. (2024). What do self-supervised speech models know about words?. *Transactions of the Association for Computational Linguistics*, *12*, 372–391. https://doi.org/10.1162/tacl_a_00656

Poli, M., Schatz, T., Dupoux, E., & Lavechin, M. (2024). Modeling the initial state of early phonetic learning in infants. *Language Development Research*, *5*(1), 1–34. https://doi.org/10.34842/y89t-6q31

Salakhutdinov, R. R., and Hinton, G. E. (2009). Deep Boltzmann machines. In Proceedings of the 12th International Conference on Artificial Intelligence and Statistics. Clearwater, Florida. https://proceedings.mlr.press/v5/salakhutdinov09a.html

Shen, G., Watkins, M., Alishahi, A., Bisazza, A., & Chrupała, G. (2024). Encoding of lexical tone in self-supervised models of spoken language. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (pp. 4250–4261). https://doi.org/10.18653/v1/2024.naacl-long.239

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, *46*(1–2), 159–216. https://doi.org/10.1016/0004-3702(90)90007-M

Smolensky, P. (1999). Grammar-based connectionist approaches to language. *Cognitive Science*, *23*(4), 589–613. https://doi.org/10.1016/S0364-0213(99)00017-8

Soderstrom, M., Mathis, D., & Smolensky, P. (2006). Abstract genomic encoding of universal grammar in optimality theory. In P. Smolensky, & G. Legendre (Eds.), *The harmonic mind, from neural computation to optimality-theoretic grammar (Volume II: Linguistic and philosophical implications)* (pp. 403–471). MIT Press.

For robust research, center values, not technology

Mark Dingemanse[a]* and Christine Cuskley[b]

[a]Centre for Language Studies, Radboud University, Netherlands and [b]Research Division, CogKnit Labs, UK
mark.dingemanse@ru.nl
chrissy@cogknit.uk

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

*Corresponding author.

doi:10.1017/S0140525X26104816, e207

Abstract

Large language models are interesting, but linguistics and cognitive science should be cautious about centering any new technology as a magic bullet. Doing so reinforces the historically "narrow focus" of linguistics identified in the target article, rather than expanding our understanding of human language. We argue that centering values instead of technology is the best way to future-proof scholarly work and to keep finding out new things about language.

Kubrick's memorable title "How I learned to stop worrying and love the bomb" evokes a hapless Major Kong who, after fiddling with electrical wiring in the bay of his plane, mounts the nuclear bomb that will set off the doomsday device, riding it to his death while hooting and waving his cowboy hat. It is a fitting choice of title for an invitation to embrace technology with abandon. But there are real choices to be made. How reproducible do we want our research to be in an era when proprietary language models are becoming an unexamined default? How complicit do we want to be in polluting our information ecology with the fallout of synthetic text? How do we make principled decisions about where to focus our efforts and how to future-proof our research?

The quality and integrity of scientific research rest on a set of core values that most scientists share and uphold. A widely used code (NCCRI, 2018) mentions *honesty* (reporting processes and sources accurately), *scrupulousness* (being precise and thoughtful at every stage of the research), *transparency* (working according to open science principles), *independence* (being impartial and not beholden to commercial or political interests), and *responsibility* (aiming to avoid environmental and societal harms). Language models present an unprecedented opportunity to undermine all of these principles in one fell swoop. Ever since BERT, their training data has been a carefully guarded and legally fraught secret (Liesenfeld, 2025; Samuelson, 2023); their frictionless design blinds us to underlying processes and transformations (Drucker, 2024); their blackboxed nature endangers reproducible research (Spirling, 2023); subscription-based access control leads to user lock-in (Widder, Whittaker, & West, 2024); and the computational burdens of training and inference put exorbitant demands on energy and environment (Luccioni, Jernite, & Strubell, 2024). Understanding human language is meant to be an endeavor to reveal the workings of what was once a black box, not an invitation to pay-per-query for access to a new one.

Dazzling new capabilities have a way of obscuring socio-technological complications (Suchman, 2019). Thousands of papers published in the period 2021–2023 relied on nominally free access to OpenAI's text-davinci-002 language model. In late 2023, OpenAI suddenly phased out this model, undercutting at one blow the reproducibility of all this work (Liesenfeld, Lopez, & Dingemanse, 2023). One affected study purported to show that "large language models are better than linguists"; its authors had to note that after text-davinci-002 was withdrawn, "the GPT 3 model that we were able to use as a replacement ... is less sophisticated at understanding complex tasks" (Ambridge & Blything, 2024). The result is a scientific edifice built on shifting sands, where "findings" have best-by dates haphazardly set by commercial parties. Blind appeal to technological advances won't suffice: for new models offered as replacements, there is typically *less* clarity about training data, fine-tuning procedures, and system prompts imposed by model providers, all of which shape input–output transformations in ways that are unobservable and irreproducible (Han et al., 2026; Zhou et al., 2024). Open-source models might provide some respite, but the most widely used models are typically open weights, meaning that we lack the data and documentation needed to understand and meaningfully manipulate model behavior (Liesenfeld & Dingemanse, 2024). When such information surfaces, it often puts supposed abilities in a new light, much like the excitement over ChatGPT passing the bar exam withers under a cold hard look at the training data and measurement methods (Martínez, 2025). Court evidence revealed that Meta's Llama, like many language models, has been trained not only on Internet-scraped data but also on the LibGen database of pirated books (Reisner, 2025), which includes thousands of works in linguistics. We would be in considerably less awe at a child effortlessly acquiring, e.g., the Russian case system if we found they had been somehow trained on virtually all linguistic descriptions of that system ever written.

Current language models are fascinating technological artifacts that deserve careful study. Classic work (Suchman, 2007; Weizenbaum, 1976) and modern research (Askari et al., 2025; Mayor, Bietti, & Bangerter, 2025; Pütz & Esposito, 2024; Smit, Smits, & Merrill, 2024) already provide constructive ways to think about the cognitive and linguistic capacities at play, going far beyond the properties of disembodied, decontextualized strings of linguistic tokens. While the target article acknowledges a "narrowness in focus" within generative linguistics (§3.3.1), we argue that language models do not expand this focus sufficiently to justify their uncritical adoption as *de facto* essential tools in understanding language. The consequences of casting language models as central to understanding human language are neither socially nor scientifically benign. Besides undermining scientific integrity, uncritical adoption risks inflating model capacities by taking illusions of understanding at face value (Messeri & Crockett, 2024) and intensifies the hyperfocus on text over talk within cognitive science (Cuskley, Woods, & Flaherty, 2024).

We reject the frame that linguists should "learn to stop worrying." Concern about the reliability and validity of models and tools is at the core of effective scientific practice. Calls for the uncritical adoption of language models ring hollow when there is a growing array of well-informed research that makes visible how language models reify and reinforce narrow views of language (Birhane & McGann, 2024; Birhane, 2021; Rasenberg et al., 2023) and how they affect the language sciences as both a community of practice and a scientific endeavor (Guest et al., 2025; Jacobs, 2023; Palmer, Smith, & Spirling, 2024; van Rooij et al., 2023).

Language models are not the first technology with revolutionary pretensions and they won't be the last (Franklin, 1999; Illich, 1973). A tech-first perspective tells us to stop worrying and embrace the inevitable. A values-first perspective asks how we can best uphold the values and standards that make our research well-rounded, robust, and reproducible. Robust research centers values over technology, moving from unexamined techno-positivism to a concern with doing the best science possible.

Financial support. MD is supported by the Vici grant *Futures of Language* (VI.C.231.103). The authors thank N. J. Enfield, Iris van Rooij, and Olivia Guest for their feedback and inspiration. Open access funding provided by Radboud University Nijmegen.

Competing interests. No competing interests to declare.

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

References

Ambridge, B., & Blything, L. (2024). Large language models are better than theoretical linguists at theoretical linguistics. *Theoretical Linguistics*, *50*(1–2), 33–48. https://doi.org/10.1515/tl-2024-2002

Askari, R., Zarrieß, S., Alacam, Ö., & Sieker, J. (2025). Are baby LMs deaf to gricean maxims? a pragmatic evaluation of sample-efficient language models. In L. Charpentier, L. Choshen, R. Cotterell, M. Omer Gul, M. Y. Hu, J. Liu, J. Jumelet, T. Linzen, A. Mueller, C. Ross, R. Sanjay Shah, A. Warstadt, E. Gotlieb Wilcox, & A. Williams (Eds.), *Proceedings of the first babyLM workshop* (pp. 52–65). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.babylm-main.4

Birhane, A. (2021). The impossibility of automating ambiguity. *Artificial Life*, *27*(1), 44–61. https://doi.org/10.1162/artl_a_00336

Birhane, A., & McGann, M. (2024). Large models of what? Mistaking engineering achievements for human linguistic agency. arXiv. https://doi.org/10.48550/arXiv.2407.08790

Cuskley, C., Woods, R., & Flaherty, M. (2024). the limitations of large language models for understanding human language and cognition. *Open Mind*, *8*, 1058–1083. https://doi.org/10.1162/opmi_a_00160

Drucker, J. (2024). Normative digital humanities. In J. O'Sullivan (Ed.), *The bloomsbury handbook to the digital humanities* (Bloomsbury Handbooks), (pp. 7–18). Paperback edition. Bloomsbury Academic.

Franklin, U. M. (1999). *The real world of technology* (CBC Massey Lectures). Revised edition. Anansi.

Guest, O., Suarez, M., Müller, B., van Meerkerk, E., Oude Groote Beverborg, A., de Haan, R., Reyes Elizondo, A., Blokpoel, M., Scharfenberg, N., Kleinherenbrink, A., Camerino, I., Woensdregt, M., Monett, D., Brown, J., Avraamidou, L., Alenda-Demoutiez, J., Hermans, F., & van Rooij. I. (2025). Against the Uncritical Adoption of "AI" Technologies in Academia. Zenodo. https://zenodo.org/records/17065099 (15 September, 2025).

Han, S., Tan, Tao., Miao, Y., Chen, X., & Sun, N. (2026). Prompting instability: An empirical study of LLM robustness in code vulnerability detection. In L. Miaomiao, Y. Xin, X. Chang, & S. Yiliao (Eds.), *AI 2025: Advances in artificial intelligence* (pp. 233–245). Springer Nature. https://doi.org/10.1007/978-981-95-4969-6_18

Illich, I. (1973). *Tools for conviviality (Open Forum)*. Calder and Boyars.

Jacobs, C. (2023). Why ChatGPT is bad for open psycholinguistics. Substack newsletter. Scidentity Crisis. https://cxjacobs.substack.com/p/why-chatgpt-is-bad-for-open-psycholinguistics. (29 January, 2024).

Liesenfeld, A. (2025). BERT: The original sin of open-washing large language models. European Open Source AI Index. https://osai-index.eu/news/bert-original-sin-openwashing

Liesenfeld, A., & Dingemanse, M. (2024). Rethinking open source generative AI: open-washing and the EU AI Act. In The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), 1774–1787. Rio de Janeiro, Brazil: ACM. https://doi.org/10.1145/3630106.3659005

Liesenfeld, A., Lopez, A., & Dingemanse, M. (2023). Opening up ChatGPT: tracking openness, transparency, and accountability in instruction-tuned text generators. In CUI '23: Proceedings of the 5th International Conference on Conversational User Interfaces. https://doi.org/10.1145/3571884.3604316

Luccioni, S., Jernite, Y., & Strubell, E. (2024). Power Hungry Processing: Watts Driving the Cost of AI Deployment? In Proceedings of the 2024 ACM Conference on Fairness, *Accountability, and Transparency* (FAccT '24), 85–99. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3630106.3658542

Martínez, E. (2025). Re-evaluating GPT-4's bar exam performance. *Artificial Intelligence and Law*, *33*(3), 581–604. https://doi.org/10.1007/s10506-024-09396-9

Mayor, E., Bietti, L. M., & Bangerter, A. (2025). Can large language models simulate spoken human conversations? *Cognitive Science*, *49*(9), e70106. https://doi.org/10.1111/cogs.70106

Messeri, L., & Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. *Nature*, *627*(8002), 49–58. https://doi.org/10.1038/s41586-024-07146-0

NCCRI. (2018). *Netherlands Code of Conduct for Research Integrity*. https://doi.org/10.17026/dans-2cj-nvwu

Palmer, A., Smith, N. A., & Spirling, A. (2024). Using proprietary language models in academic research requires explicit justification. *Nature Computational Science*, *4*(1), 2–3. https://doi.org/10.1038/s43588-023-00585-1

Pütz, O., & Esposito, E. (2024). Performance without understanding: How ChatGPT relies on humans to repair conversational trouble. *Discourse & Communication*, *18*(6), 859–868. https://doi.org/10.1177/17504813241271492

Rasenberg, M., Amha, A., van Koppen, M., van Miltenburg, E., de Rijk, L., Stommel, W., & Dingemanse, M. (2023). Reimagining language: Towards a better understanding of language by including our interactions with non-humans. *Linguistics in the Netherlands*, *40*, 309–317. https://doi.org/10.1075/avt.00095.ras

Reisner, A. (2025). The Unbelievable Scale of AI's Pirated-Books Problem. The Atlantic. https://www.theatlantic.com/technology/archive/2025/03/libgen-meta-openai/682093/. (16 April, 2025).

Samuelson, P. (2023). Generative AI meets copyright. *Science*, *381*(6654), 158–161. https://doi.org/10.1126/science.adi0656

Smit, R., Smits, T., & Merrill, S. (2024). Stochastic remembering and distributed mnemonic agency: Recalling twentieth century activists with ChatGPT. *Memory Studies Review*, *1*(aop), 1–22. https://doi.org/10.1163/29498902-202400015

Spirling, A. (2023). Why open-source generative AI models are an ethical way forward for science. *Nature*, *616*(7957), 413–413. https://doi.org/10.1038/d41586-023-01295-4

Suchman, L. A. (2007). *Human-machine reconfigurations: Plans and situated actions*. (2nd ed.) Cambridge University Press.

Suchman, L. A. (2019). Demystifying the intelligent machine. In H. Teresa (Ed.), *Cyborg futures: Cross-disciplinary perspectives on artificial intelligence and robotics* (Social and Cultural Studies of Robots and AI). (pp. 35–61). Springer International Publishing. https://doi.org/10.1007/978-3-030-21836-10_3

van Rooij, I., Guest, O., Adolfi, F. G., de Haan, R., Kolokolova, A., & Rich, P. (2023). Reclaiming AI as a theoretical tool for cognitive science. *Computational Brain & Behavior 7*(4), 616-636. https://doi.org/10.1007/s42113-024-00217-5

Weizenbaum, J. (1976). *Computer power and human reason: from judgment to calculation*. W. H. Freeman.

Widder, D. G., Whittaker, M., & Myers West, S. (2024). Why 'open' AI systems are actually closed, and why this matters. *Nature*, *635*(8040), 827–833. https://doi.org/10.1038/s41586-024-08141-1

Zhou, L., Schellaert, W., Fernando Martínez-Plumed, Y. M., Ferri, C., & Hernández-Orallo, J. 2024. Larger and more instructable language models become less reliable. *Nature*, *634*(8032). 61–68. https://doi.org/10.1038/s41586-024-07930-y

What ever happened to meaning?

Michael Eric Goodale[a,b] and Salvador Mascarenhas[a,b]*

[a]Department of Cognitive Studies, Ecole Normale Superieure, Paris, France and [b]Institut Jean-Nicod, Paris, France

salvador.mascarenhas@ens.fr

michael.goodale@ens.fr

*Corresponding author.

doi:10.1017/S0140525X2610466X, e208

Abstract

By "linguistics," the target article means usage-based linguistics, which, we agree, plays very well with language models. But the article summarily dismisses important work on *meaning* from the now disreputable tradition of generative linguistics. We refute the authors' arguments from distributional semantics and "green

tea," and highlight the importance of formal compositional semantics for the study of thought.

The target article contends that a convergence has occurred between linguistics and language models. Reading the piece, however, reveals that what is meant by "linguistics" is usage-based linguistics. Such a convergence should be surprising to no one: whichever strand of usage-based linguistics one has in mind, surely nothing could be more usage-based than the linguistic capacities of state-of-the-art language models trained on massive corpora of language in use, with no language-specific biases.

A more interesting question is whether "generative linguistics" still has any relevance today. Futrell and Mahowald say yes, absolutely, if one is interested in studying programming languages. We submit instead that generative linguistics, despite its current lack of popularity in cognitive science, actually has a great deal to offer, and is brushed aside, as the authors do, very much to the detriment of the field. Here, we focus exclusively on formal semantics, by which we mean the kind of mathematically sophisticated compositional semantics pursued in mainstream generative-linguistics departments. In short, what they're doing over at MIT.

Despite the centrality of meaning to language, the authors only spare a few words on the topic: two paragraphs on the meanings of words and one on the principle of compositionality. The authors promote distributional semantics as the right theory of the meanings of content words, and directly contrast it with the mainstream in formal semantics, citing Heim and Kratzer's (1998) textbook. But this contrast is not right. Formal semantics has never had much to say about the meanings of content words or the concepts that lie behind them. This is not an oversight, but an extremely wise methodological move by formal semantics. One could not tell from reading the target article, but the question of what *concepts* are is an unmitigated quagmire. Reams of paper have been written on the many competing theories of concepts and word meanings: essential descriptions, sensory definitions, theory-theory, prototypes and stereotypes, conceptual roles, extensions. Evidently, the authors believe that distributional semantics and vector spaces somehow definitively settle all of these fraught debates, but they do not tell us how.

Compare with the approach from formal semantics. Word meanings are concepts. Nobody knows what on earth concepts are, but we are certain that the concept LION needs to give us some sort of procedure for telling whether something is or is not a lion. So, we put a placeholder for that in our semantics, which *you* can fill in with your favorite theory of concepts. *We* will focus on combinatorial properties.

Which brings us to compositionality. The authors' paragraph addresses no question of import. Instead, it focuses on the supposedly damning problem of "green tea." Here is what we take to be the authors' concern. "Green tea" doesn't just mean tea that is green, yet it is not a completely opaque idiom. Therefore, compositionality is gradable, and if this is true of most complex expressions, then language is not all that compositional.

First, both "green tea" and full-on idioms like "kick the bucket" have perfectly compositional, literal interpretations, which can be quite useful, for example, in gallows humor or to refer to tea that is green. Moreover, just because something *seems* gradable, it does not follow that the best way to think about *it* is in gradable terms. Truth can feel quite gradable, as manifested in expressions like "This is so true," or "This is far more true than that." Yet, the mainstream view in science, as in philosophy, is that truth is *not at all* gradable, but we have notions like *probability* and *verisimilitude*, which are in fact gradable and are connected to truth. Once we recognize that each phrase has a compositional interpretation *and* potentially a host of idiomatic uses, we can say: compositionality itself is non-gradable as the mathematicians propose, but for any expression that *also* has an idiomatic use, we can consider the *similarity* between the idiomatic meaning and the compositional meaning. We do not see how it matters what precise proportion of complex expressions *also* have idiomatic readings. And by the way, what proportion out of *what*, the infinity of possible complex expressions? At any rate: problem solved, as far as we can see, with just a little charity and common sense, and without breaking faith with mathematics.

There is in fact a wealth of work in the formal-semantics literature on this and far thornier cases. Even Montague (1970/1974), in one of the founding documents of the field, addresses some of these questions. For example, a "fake lawyer" is not a lawyer who is fake (see Guerrini, 2024, for a recent treatment), nor is a "big flea" necessarily a "big creature."

Most importantly, there is a deep connection between how meanings are composed in natural language and how human thought works. A recent issue of this very journal demonstrated quite clearly that the language-of-thought hypothesis is a perfectly respectable and viable research program (Quilty-Dunn, Porot, & Mandelbaum, 2023). The idea is to explain the nature of human thought by means of discrete symbolic representations, which can be combined algebraically, something diametrically opposed to the approach recommended by Futrell and Mahowald.

Given that natural language does such a remarkable job of expressing the content of so many of our thoughts, it makes sense to wonder whether insights from formal natural-language semantics might bear on thought. Thankfully, there is a fair amount of work along these lines already, for example, providing theories of modal representations (Phillips & Kratzer, 2024) or of reasoning and decision-making (Koralus & Mascarenhas, 2013). This scholarship repurposes theories from formal semantics, which are altogether inaccessible under a view that focuses exclusively on usage patterns and statistical distributions.

Futrell and Mahowald point out that "linguistics needs the rest of science." While this is certainly true, the rest of science also needs linguistics, and not just the very particular brand of usage-based construction grammar which the authors favor.

Acknowledgments. We thank Baptiste Loreau-Unger and Benjamin Spector for helpful comments.

Financial support. This research was supported in part by a grant from École Doctorale Frontières de l'Innovation en Recherche et Education – Programme Bettencourt. Ecole Normale Supérieure-PSL's Département d'Etudes Cognitives is supported by ANR grants, ANR-10-IDEX-0001-02 and FrontCog, ANR-17-EURE-0017.

Competing interests. We have no competing interests.

References

Guerrini, J. (2024). Keeping fake simple. *Journal of Semantics*, *41*(2), 175–210. https://doi.org/10.1093/jos/ffae010

Heim, I., & Kratzer, A. (1998). *Semantics in generative grammar*. Blackwell.

Koralus, P., & Mascarenhas, S. (2013). The erotetic theory of reasoning: Bridges between formal semantics and the psychology of deductive inference. *Philosophical Perspectives*, *27*, 312–365. https://doi.org/10.1111/phpe.12029

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

Montague, R. (1974). English as a formal language. In R. H. Thomason (Ed.), *Selected papers of Richard Montague* (pp. 188–221). Yale University Press. (Original work published 1970).
Phillips, J., & Kratzer, A. (2024). Decomposing modal thought. *Psychological Review*, *131*(4), 966–992. https://doi.org/10.1037/rev0000481
Quilty-Dunn, J., Porot, N., & Mandelbaum, E. (2023). The best game in town: The re-emergence of the language of thought hypothesis across the cognitive sciences. *Behavioral and Brain Sciences*, *46*(e261). https://doi.org/10.1017/S0140525X22002849

You can't be neutral on a moving train

Jeffrey Heinz[a] and Jonathan Rawski[b*]

[a]Stony Brook University, USA and [b]San Jose State University, USA
jeffrey.heinz@stonybrook.edu
jon.rawski@sjsu.edu

*Corresponding author.

doi:10.1017/S0140525X26104798, e209

Abstract

The commentary argues the authors employ misdirection and strawmanning to cast others as polarized extremes and themselves as the reasonable centrists. We argue that these patterns of misrepresentation ultimately damage any consensus and middle ground they claim to hope to reach.

Futrell & Mahowald (F&M) want readers to see them as reasonable centrists, but do so by casting other positions as polarized extremes via misdirection and strawmen. Their centrism is an illusion and comes at the cost of saying anything substantive or advancing the conversation. By the end, the sheep becomes a wolf: Claudette is inviting Norm for dinner, and Norm is on the menu. With an invitation like that, no wonder linguists would rather continue to focus on the human capacity for language in all its glorious variations.

F&M frame radical positions as a middle ground by strawmanning the debate. For example, regarding learnability constraints, they state *"There is no logical necessity for a restrictive formalism to explain language learning."* and *"The key mistake was . . . that inductive bias has to come from restrictions on the learner's hypothesis space."* They highlight Wilson's (2025) position paper describing several deep learning phenomena using the PAC-Bayes framework via soft' biases. In reality, these are a set of hidden "hard" biases: a fixed computable hypothesis class, a fixed description-length prior tied to the architecture, I.I.D. sampling assumptions, bounded-loss assumptions, task-dependent representational primitives (by fallaciously appealing to Li and Vitanyi's (2008) "additive-constant" invariance theorem), and other capacity constraints. While description-length priors are coherent, Wolpert (2021) shows a bias only justifies a particular hypothesis class to the extent it matches the (unknown) distribution over tasks (see James et al., 2023 and Curth, Jeffares, & van der Schaar, 2023 on task-dependent complexity and double-descent). Without explicitly formalizing that alignment, moving from "real-world language data is inherently compressible" to endorsing biases like those above is meaningless. Whether hypotheses are restricted, admissible data presentations are restricted, or learning resources in time and space are unrestricted, there are fundamental tradeoffs in solving learning problems (Angluin, 1988; Goyal & Bengio, 2022, a.o.).

F&M's own candidate constraints like sensitivity and information locality are also model-dependent (Hahn & Rofin, 2024; Someya et al., 2025), task-dependent (Hahn, Jurafsky, & Futrell, 2021), and ultimately "hard." Even Wilson and colleagues agree that *"it is largely the design of the architecture that makes generalizable solutions more easily accessible"* (Goldblum et al., 2024). Like all "domain-general" biases, F&M's candidates turn out to be just "domain-specific" biases of a different sort (Rawski & Heinz, 2019). In other words, the domains and biases F&M want are real, and the others are not. F&M can only achieve centrism from such an extreme position by misassigning people to trivially false claims, like that Rawski and Baumont (2023)'s 1-page paper says Large Language Models (LLMs) *"can learn to approximate anything"* when it does not. Strawmanning debates by mischaracterizing one perspective to position oneself as a middle ground, while ignoring the substance of the issues, is disingenuous and ultimately leads nowhere.

F&M cover these realities with doublespeak and fallacious arguments. They hype LLM success, concede the successes are disagreed upon and may always emerge from shallow heuristics (Liu et al., 2022; McCoy, Pavlick, & Linzen, 2019; Petty et al., 2025; Vafa et al., 2025; Vázquez Martínez et al., 2026, a.o.), backtrack to *"this is not the right place to adjudicate these issues,"* and then reassert that LLMs have *"learned 'the real thing'"* and are *"not purely relying on heuristics or memorization or cheap tricks."* Even their interpretability "thought experiment" about an idealized "c-command neuron" is a revamp of standard cognitive neuroscience techniques with the same reification fallacies and circularity (Buzsaki, 2019). Again they admit *"LMs just don't represent anything in such human-interpretable ways,"* yet insist they are *"linguistically interesting."* This is not a serious discussion.

Beyond the banal observation that *"there are increasingly researchers using language models to ask questions that are informed by and which can inform linguistic theory,"* what is F&M's actual point? Simultaneously insisting on the Eliminativist fallacy (Smolensky, 1988) that restrictive symbolic formalisms are *"dogma,"* while claiming *"we are not eliminativists about grammatical structure,"* is incoherent. Taking a Limitivist view (Smolensky, 1988) where symbolic formalisms are an "approximate explanation" and *"graded, probabilistic, and fuzzy"* subsymbolic primitives are cognitively real, they bizarrely omit 30 years of neural networks in generative linguistics via Optimality Theory and Harmonic Grammar (Prince & Smolensky, 1997; Smolensky & Legendre, 2006). Advocating for "domain-general" learning biases, which end up being domain-specific, paints them into a Dennettian corner to conclude the "troll's truism" (Schackel, 2005) that *"linguistic structure is real."* Little else remains, certainly nothing new.

Here we see the tragedy of the centrist, ruled by *"the zeitgeist"* in F&M's terms, at the cost of ideas or content. Like most self-described centrists, however, their strategies serve to cover another goal. F&M seem to want linguistics to effectively become a niche within psychology (Núñez et al., 2019), where replicability and theory crises (Eronen & Bringmann, 2021) mean *"theories rise and decline, come and go, more as a function of baffled boredom than anything else; and the enterprise shows a disturbing absence of that cumulative character that is so impressive"* (Meehl, 1978). Linguistic theory is so robust and successful at phenomena discovery that LLM performance benchmarks (Gauthier et al., 2020, a.o.) and evaluations (Hale & Stanojević, 2024, a.o.) crucially rely

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

on them. When neural models are used to trivialize the substance of linguistic theory, as F&M do with competence and performance (Berwick, 2018; Firestone, 2020; West et al., 2024), linguists have no reason to engage. Replacing theory-driven science with model-driven science, and phenomena with effects-chasing, is a profound retreat from the core insights which gained linguistics membership in cognitive science in the first place.

F&M wonder why many linguists remain uninterested in LLMs, which have not yet discovered virtually any new phenomena nor explained anything of theoretical significance. Yes, LLMs are a technological advance and engineering achievement. Nonetheless, language models, in whatever guise, will continue to be a scientific embarrassment to language scientists as long as they do not offer robust phenomena discovery, counterfactual support, explanatory depth, and other scientific insight into the fundamental properties of the human capacity for language.

Financial support. This work is supported by NSF Award #2416183 to JH.

Competing interests. None.

References

Angluin, D. (1988). Identifying languages from stochastic examples (p. 19). Technical Report 614. Yale University. Department of Computer Science.

Berwick, R. (2018). Revolutionary New Ideas Appear Infrequently. In *Syntactic Structures after 60 Years*. Norbert Hornstein, Howard Lasnik, Pritty Patel-Grosz and Charles Yang, eds. de Gruter Mouton.

Buzsáki, G. (2019). *The brain from inside out*. Oxford University Press.

Curth, A., Jeffares, A., & van der Schaar, M. (2023). A u-turn on double descent: Rethinking parameter counting in statistical learning. *Advances in Neural Information Processing Systems*, *36*, 55932–55962.

Eronen, M. I., & Bringmann, L. F. (2021). The theory crisis in psychology: How to move forward. *Perspectives on Psychological Science*, *16*(4), 779–788. https://doi.org/10.1177/1745691620970586

Firestone, C. (2020). Performance vs. Competence in human–machine comparisons. *Proceedings of the National Academy of Sciences*, *117*(43), 26562–26571.

Gauthier, J., Hu, J., Wilcox, E., Qian, P., & Levy, R. (2020). SyntaxGym: An Online Platform for Targeted Evaluation of Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76, Online. Association for Computational Linguistics.

Goldblum, M., Finzi, M. A., Rowan, K., & Wilson, A. G. (2024). Position: the no free lunch theorem, Kolmogorov complexity, and the role of inductive biases in machine learning. In *Forty-first International Conference on Machine Learning*.

Goyal, A., & Bengio, Y. (2022). Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *478*(2266), 20210068.

Hahn, M., Jurafsky, D., & Futrell, R. (2021). Sensitivity as a complexity measure for sequence classification tasks. *Transactions of the Association for Computational Linguistics*, *9*, 891–908.

Hahn, M., & Rofin, M. (2024). Why are Sensitive Functions Hard for Transformers?. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers) (pp. 14973–15008).

Hale, J., & Stanojević, M. (2024). Do LLMs learn a true syntactic universal?. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17106–17119, Miami, Florida, USA. Association for Computational Linguistics.

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An introduction to statistical learning: With applications in Python*. Springer International Publishing.

Li, M., & Vitányi, P. (2008). *An introduction to Kolmogorov complexity and its applications*. Springer.

Liu, B., Ash, J. T., Goel, S., Krishnamurthy, A., & Zhang, C. (2022).Transformers learn shortcuts to automata. In *The Eleventh International Conference on Learning Representations*.

McCoy, R. T., Pavlick, E., & Linzen, T. (2019). Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3428–3448).

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*(4), 806–834. https://doi.org/10.1037/0022-006X.46.4.806

Núñez, R., Allen, M., Gao, R., Miller Rigoli, C., Relaford-Doyle, J., & Semenuks, A. (2019). What happened to cognitive science?. *Nature Human Behaviour*, *3*(8), 782–791.

Petty, J., Hu, M., Wang, W., Ravfogel, S., Merrill, W., & Linzen, T. (2025). RELIC: Evaluating compositional instruction following via language recognition. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks*, Emergent Abilities, and Scaling.

Prince, A., & Smolensky, P. (1997). Optimality: From neural networks to universal grammar. *Science*, *275*(5306), 1604–1610.

Rawski, J., & Baumont, L. (2023). Modern language models refute nothing. Lingbuzz Preprint: https://ling.auf.net/lingbuzz/007203

Rawski, J., & Heinz, J. (2019). No free lunch in linguistics or machine learning: Response to Pater. *Language*, *95*(1), e125–e135.

Shackel, N. (2005). The vacuity of postmodernist methodology. *Metaphilosophy*, *36*(3), 295–320.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, *11*(1), 1–23.

Smolensky, P., & Legendre, G. (2006). *The harmonic mind: From neural computation to optimality-theoretic grammar. Volumes I and II*. MIT Press.

Someya, T., Svete, A., DuSell, B., O'Donnell, T., Giulianelli, M., & Cotterell, R. (2025). Information Locality as an Inductive Bias for Neural Language Models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27995–28013, Vienna, Austria. Association for Computational Linguistics.

Vafa, K., Chang, P. G., Rambachan, A., & Mullainathan, S. (2025). What Has a Foundation Model Found? Using Inductive Bias to Probe for World Models. In *International Conference on Machine Learning* (pp. 60727–60747).

Vázquez Martínez, H. J., Heuser, A., Yang, C., & Kodner, J. (2026). Evaluating the Existence Proof: LLMs as Cognitive Models of Language Acquisition. In José-Luis Mendívil-Giró (Ed.), *Artificial Knowledge of Language*. Vernon Press.

West, P., Lu, X., Dziri, N., Brahman, F., Li, L., Hwang, J. D., Jiang, L., Fisher, J., Ravichander, A., Chando, K., Newman, B., Koh, P.W., Ettimner, A., & Choi, Y. (2024). The Generative AI Paradox: "What It Can Create, It May Not Understand". In *The Twelfth International Conference on Learning Representations*.

Wilson, A. G. (2025). Position: Deep Learning is Not So Mysterious or Different. In *International Conference on Machine Learning* (pp. 82326–82346). PMLR.

Wolpert, D. H. (2021). What is important about the no free lunch theorems?. In Black box optimization, machine learning, and no-free lunch theorems (pp. 373–388). Cham: Springer International Publishing.

LLMs are not children: They have to earn our love

Jeffrey Lidz[a]* and Katherine Howitt[a]

[a]Department of Linguistics, University of Maryland at College Park, College Park, MD, USA
jlidz@umd.edu
kghowitt@umd.edu

doi:10.1017/S0140525X26104695, e210

Abstract

Futrell and Mahowald argue that the success of large language models should move the field away from the formal structures of generative linguistic theory. The limited success of these models falls short of formal linguistic theory in explaining both the character of human languages and understanding the trajectory of child language acquisition.

Futrell and Mahowald (F&M) present themselves as offering a measured middle ground between the extreme claims that large language models (LLMs) obviate the need for linguistic theory and the (in their estimation) equally extreme claim that humans possess language-specific innate inductive bias. However, despite this posture, they align themselves aggressively with the former, without fully engaging the empirical discoveries that have propelled the latter forward. Here, we limit our discussion to the structure of filler-gap dependencies and their emergence during language development.

F&M present these models as the central path forward for linguistic theory. But despite their optimism, LLMs are simply not as good at language as they claim. LLMs appear to succeed at recognizing not only simple filler-gap dependencies but also the environments in which they cannot occur – syntactic islands (Wilcox et al., 2018, 2024). Islands represent prototypical evidence for innate bias in generative linguistics because the patterns are semantically unpredictable and share structure only at a high level of abstraction (Boeckx, 2008; Chomsky, 1977; Ross, 1967), so LLM success matters. However, while LLMs discover island sensitivity for embedded *wh*-movement (Wilcox et al., 2024), their behavior is less robust than that found in human speakers (Sprouse 2007; Sprouse et al., 2016, i.a.). Yet despite these limitations, LLMs' successes are touted as powerful evidence against nativist theorizing: islands do not require innate bias. This claim misses the broader observation that motivates islands' centrality: islands do not just constrain *wh*-questions (1), but also surface-dissimilar dependencies such as clefting (2), topicalization (3), tough movement (4), and comparative VP-ellipsis (5), as well as in filler-gap dependencies across unrelated languages.

(1) * What did the engineers ignore the claim that humans need?
(2) * It is syntax that engineers ignore the claim that humans need.
(3) * Syntax, engineers ignore the claim that humans need.
(4) * Syntax is easy for the engineers to ignore the claim that humans need.
(5) * Engineers need less syntactic bias than I believe the claim that humans do.

Syntacticians posit a representational format for filler-gap dependencies that gives rise to parallel behavior across phenomena both within and across languages, including *wh*-agreement in Irish (McCloskey, 2001), *wh*-agreement in Chamorro (Chung, 2009), quantifier float in Irish-English (McCloskey, 2000), scope islands in Mandarin (Huang, 1982), comparatives in Greek (Merchant, 2009), the interaction of scope, binding, and movement in English (Fox, 1999), and even in children's nontarget productions in English (Liter, Grolla & Lidz, 2022; Lutken, Legendre, & Omaki, 2020; Thornton, 1990), Brazilian Portuguese (Grolla, 2022), and French (Oiry & Demirdache, 2006).

LLMs instantiate an alternative: constraints on filler-gap dependencies can be extracted from the surface distributions of strings in the language. However, to cover the same range of phenomena as humans (i.e., (1)–(5)), they must either access enough data to arrive at consistent hypotheses individually for each construction, or recognize a fundamental similarity across constructions based on surface distributions, and then use the discovered island sensitivity in one construction to generalize broadly.

Whether human children see enough data to arrive at such hypotheses independently is an open empirical question, but LLMs, given magnitudes more input than human children, do not (Howitt et al., 2024; Ozaki et al., 2022). Further, LLMs extract *contradictory* hypotheses where evidence for one construction's island sensitivity is taken as evidence against another's. Howitt et al. (2024) showed that training an LLM on grammatical instances of clefting (e.g., *It was syntax that the humans needed*) allowed the model to correctly posit islands for clefts, but also led them to lose island sensitivity in *tough* constructions (4). And while some work suggests that LLMs sometimes use the same representational networks to inform their representations of some filler-gap dependencies (Boguraev et al., 2025), these findings and, in fact, the methods that reveal them are dependent on the highly similar surface forms of this subset of constructions.

While F&M's research program invites us to expand our search for inductive biases away from grammatical architecture, the support for this invitation is based on hope, not evidence. LLMs might capture surprising distributional facts in one language, but in the absence of architectural constraints, observed cross-linguistic and cross-constructional consistency would have to depend on an extralinguistic source. Critics of an architectural explanation point to communicative pressures as this extralinguistic source (Abeillé et al., 2020; Cuneo & Goldberg, 2023; Erteschik-Shir, 1973; Goldberg, 2006), but given the range of phenomena that display island sensitivity and the diversity of the uses to which these phenomena are put, such hypotheses fall short of the explanatory mark (Lidz & Williams, 2009; Cartner et al., 2025).

Finally, F&M's research program, like many classic works (Chomsky, 1975; Gold, 1967; Wexler & Culicover, 1980), addresses learnability abstractly but provides little insight into how knowledge grows. We now know that local argument structure dependencies are acquired by 15–16 months (Fisher, Jin, & Scott, 2020; Lidz, White, & Baier, 2017; Perkins & Lidz, 2019), a few months prior to nonlocal dependencies in *wh*-questions (Perkins & Lidz, 2021). Explicit, interpretable models of language acquisition reveal (a) how learners could actively select a subset of data to learn from to arrive at argument structure representations first (Perkins, Feldman, & Lidz, 2022) and (b) that this knowledge is necessary to identify the form of filler-gap dependencies (Perkins, Feldman, & Lidz, 2025). These models of acquisition allow us to explain not just acquired knowledge but specific developmental steps that learners take to acquire that knowledge. LLMs, by their very nature, do not afford developmental analysis.

It is no surprise that an LLM trained on 450 million years of input has a pretty good idea about what is likely to occur in a language. But an 18-month-old with just 450 days of language experience also has a pretty good idea, and we understand both the steps they take to get there and the reasons those steps occur in the order that they do. The partial successes of LLMs are impressive. But our goal should not be to make these models better, but to

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

make better models that respect the empirical landscape identified by theoretical and developmental linguists.

Financial support. This work was supported in part by NSF BCS-2336014 to JL.

Competing interests. The authors have no competing interests.

References

Abeillé, A., Hemforth, B., Winckel, E., & Gibson, E. (2020). Extraction from subjects: Differences in acceptability depend on the discourse function of the construction. *Cognition*, *204*, 104293. https://doi.org/10.1016/j.cognition.2020.104293

Boeckx, C. (2008). Islands. *Language and Linguistics Compass*, *2*(1), 151–167. https://doi.org/10.1111/j.1749-818X.2007.00043.x

Boguraev, S., Potts, C., & Mahowald, K. (2025). Causal interventions reveal shared structure across English filler–gap constructions. In C. Christodoulopoulos, T. Chakraborty, C. Rose, & V. Peng (Eds.), *Proceedings of the 2025 conference on empirical methods in natural language processing* (pp. 25032–25053). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.emnlp-main.1271

Cartner, M., Kogan, M., Webster, N., Wagers, M. & Sichel, I. (2025). Subject islands do not reduce to construction-specific discourse function. Cognition. Article in press.

Chung, S. (2009). Six arguments for wh-movement in Chamorro. In D. B. Gerdts, J. C. Moore, & M. Polinsky (Eds.), *Hypothesis A/Hypothesis B: Linguistic explorations in honor of David M. Perlmutter* (pp. 91–110). The MIT Press.

Chomsky, N. (1975). *Reflections on language*. Pantheon Books.

Chomsky, N. (1977). On wh-movement. In P. Culicover, A. Akmajian, & T. Wasow (Eds.), *Formal syntax*. (pp. 71–133). Academic Press.

Cuneo, N., & Goldberg, A. E. (2023). The discourse functions of grammatical constructions explain an enduring syntactic puzzle. *Cognition*, *240*, 105563. https://doi.org/10.1016/j.cognition.2023.105563

Erteschik-Shir, N. (1973). On the nature of island constraints (Unpublished doctoral dissertation). MIT.

Fisher, C., Jin, K., & Scott, R. (2020) The developmental origins of syntactic bootstrapping. *Topics in Cognitive Science*, *12*(1), 48–77.

Fox, D. (1999). Reconstruction, binding theory, and the interpretation of chains. *Linguistic Inquiry*, *30*(2), 157–196.

Gold, E. M. (1967). Language identification in the limit. *Information and Control*, *10*(5), 447–474. https://doi.org/10.1016/S0019-9958(67)91165-5

Goldberg, A. (2006). *Constructions at work: The nature of generalization in language*. Oxford University Press.

Grolla, E. (2022). Syntactic constraints and medial wh-questions in child Brazilian Portuguese. In Y. Gong, & F. Kpogo (Eds.), *Proceedings of the 46th Annual Boston University conference on language development* (pp. 256–269). Cascadilla Press.

Howitt, K., Nair, S., Dods, A., & Hopkins, R. M. (2024). Generalizations across filler-gap dependencies in neural language models. In L. Barak, & M. Alikhani (Eds.), *Proceedings of the 28th conference on computational natural language learning* (pp. 269–279). Association for Computational Linguistics. https://aclanthology.org/2024.conll-1.21

Huang, C. T. J. (1982). Logical relations in Chinese and the theory of grammar (Unpublished doctoral dissertation). MIT.

Lidz, J., White, A. S., & Baier, R. (2017). The role of incremental parsing in syntactically conditioned word learning. *Cognitive Psychology*, *97*, 62–78. https://doi.org/10.1016/j.cogpsych.2017.06.002

Lidz, J. & Williams, A. (2009). Constructions on holiday. *Cognitive Linguistics*, *20*(1), 177–189. https://doi.org/10.1515/COGL.2009.011

Liter, A., Grolla, E., & Lidz, J. (2022). Cognitive inhibition explains children's production of medial wh-phrases. *Language Acquisition*, *29*(3), 327–359. https://doi.org/10.1080/10489223.2021.2023813

Lutken, C. J., Legendre, G., & Omaki, A. (2020). Syntactic creativity errors in children's wh-questions. *Cognitive Science*, *44*(7), e12849. https://doi.org/10.1111/cogs.12849

McCloskey, J. (2000). Quantifier float and wh-movement in Irish English. *Linguistic Inquiry*, *31*(1), 57–84. https://doi.org/10.1162/002438900554299

McCloskey, J. (2001). The morphosyntax of wh-extraction in Irish. *Journal of Linguistics*, *37*(1), 67–100. http://www.jstor.org/stable/4176643

Merchant, J. (2009). Phrasal and clausal comparatives in Greek and the abstractness of syntax. *Journal of Greek Linguistics*, *9*(1), 134–164. https://doi.org/10.1163/156658409X12500896406005

Oiry, M., & Demirdache, H. (2006). Evidence from L1 acquisition for the syntax of wh-scope marking in French. In V. Torrens, & L. Escobar (Eds.), *The acquisition of syntax in romance languages* (pp. 289–315). John Benjamins.

Ozaki, S., Yurovsky, D., & Levin, L. (2022). How well do LSTM language models learn filler-gap dependencies? In A. Ettinger, T. Hunter, & B. Prickett (Eds.), *Proceedings of the society for computation in linguistics* (pp. 76–88). online: Association for Computational Linguistics. https://aclanthology.org/2022.scil-1.6/

Perkins, L., Feldman, N., & Lidz, J. (2022). The power of ignoring: Filtering input for argument-structure acquisition. *Cognitive Science*, *46*(1), e13080. https://doi.org/10.1111/cogs.13080.

Perkins, L., Feldman, N., & Lidz, J. (2025). Mind the Gap: Learning the surface forms of movement dependencies. Language. Article in press.

Perkins, L. & J. Lidz (2019). Filler-gap dependency comprehension at 15-months: The role of vocabulary. *Language Acquisition*, *27*(1), 98–115.

Perkins, L. & J. Lidz (2021). Eighteen-month-old infants represent nonlocal syntactic dependencies. *Proceedings of the National Academy of Sciences*, *118*(41), e2026469118. https://doi.org/10.1073/pnas.2026469118.

Ross, J. R. (1967). Constraints on variables in syntax. (Doctoral dissertation, Massachusetts Institute of Technology).

Sprouse, J. (2007). *A program for experimental syntax*. University of Maryland dissertation.

Sprouse, J., Caponigro, I., Greco, C., & Cecchetto, C. (2016). Experimental syntax and the variation of island effects in English and Italian. *Natural Language & Linguistic Theory*, *34*(1), 307–344.

Thornton, R. (1990). Adventures in long-distance moving: The acquisition of complex wh-questions (Doctoral dissertation). University of Connecticut, Storrs, CT.

Wexler, K. & P. Culicover (1980). *Formal principles of language acquisition*. MIT Press.

Wilcox, E., Futrell, R., & Levy, R. (2024). Using computational models to test syntactic learnability. *Linguistic Inquiry*, *55*, 1–44. https://doi.org/10.1162/ling_a_00491

Wilcox, E., Levy, R., Morita, T., & Futrell, R. (2018). What do RNN language models learn about filler–gap dependencies? In T. Linzen, G. Chrupała, & A. Alishahi (Eds.), *Proceedings of the 2018 EMNLP workshop blackboxNLP: Analyzing and interpreting neural networks for NLP* (pp. 211–221). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-5423

Language models as tools for investigating the distinction between possible and impossible natural languages

Julie Kallini*† and Christopher Potts*†

Stanford University, Stanford, CA, USA
kallini@stanford.edu
cgpotts@stanford.edu

*Corresponding author.
†Equal contribution. Author order determined alphabetically by last name.

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

doi:10.1017/S0140525X26104737, e211

Abstract

We argue that language models (LMs) have strong potential as investigative tools for probing the distinction between possible and impossible natural languages and thus uncovering the inductive biases that support human language learning. We outline a phased research program in which LM architectures are iteratively refined to better discriminate between possible and impossible languages, supporting linking hypotheses to human cognition.

Which conceivable linguistic systems are possible for humans to learn and use as natural languages? A complete answer to this question would yield profound insights into the human capacity for language. However, our tools for addressing the question are very limited. The historical record is shaped by external factors (e.g., patterns of migration and contact). Artificial language learning experiments happen in highly limited settings and cannot avoid influences from participants' own languages, and any attempt to expose children only to impossible languages would be deeply unethical.

A number of recent papers present evidence that language models (LMs) learn possible human languages more efficiently than impossible ones. In light of how limited our existing toolkit is, this sounds like a promising development. Futrell and Mahowald interpret these studies as evidence that LMs possess inductive biases aligned with human languages, such as a preference for information locality.

What precisely is the value of such findings for the project of trying to understand what makes a language possible? On the face of it, the answer is not immediately clear. After all, LMs differ from humans in the data they consume, the data they produce, and how they learn. This might seem to make evidence derived from LMs irrelevant by definition.

Our own view (consistent with Futrell and Mahowald's perspective) is that LMs have great potential as investigative tools for understanding the possible/impossible distinction. We present our argument as a series of research phases:

Phase 1: Find examples of possible and impossible languages. We might stipulate that any attested language is a possible language. For impossible languages, we assume there are clear cases that are very distant from the attested ones (e.g., languages that have no predictable structure). Linguists have also posited relatively clear instances of impossible languages that are closer to the attested ones (Moro, Greco, & Cappa, 2023; Moro, 2008), and these might be the most informative ones in terms of theory building.

Phase 2: Train LMs on minimal pairs of possible and impossible languages from Phase 1. Mitchell and Bowers (2020) helped begin this project. Kallini et al. (2024) present evidence that such models learn English more efficiently than counterfactual impossible languages. Xu et al. (2026) and Yang et al. (2025) expand the empirical scope of these claims and arrive at similar conclusions. Ziv et al. (2026) present a more mixed picture. Instances of alignment and misalignment are both valuable in this context.

Phase 3: Using the evidence gathered in Phase 2, study the inductive biases of these LMs and use these insights to inform statements about what the corresponding human biases are. Crucially, this will require us to state rigorous linking hypotheses between LM constructs and human cognition. The real challenge lies in stating linking hypotheses that are informative. Precedents from neuroscience make us optimistic about this project (e.g., McIntosh et al., 2016).

Phase 4: Explore novel LM architectures, training datasets, and training objectives, seeking designs that are even more successful at distinguishing the possible from the impossible than those used in Phase 3. We see many opportunities. For example, many present-day LMs have mechanisms that give them what is, in effect, perfect memory over long sequences of words. By reducing the capacity of these mechanisms, we might encourage even more locality and thus better match human language. Implicit in this description is a hypothesis linking LM memory to human memory.

Once Phase 4 is reached, we return to Phase 2 for empirical evaluation against the languages we found in Phase 1. This leads to new linking hypotheses in Phase 3 and, in turn, to new LM innovations in Phase 4.

On the above approach, we do not immediately get insights into which languages are possible and which are impossible (Phase 1). However, as we gain confidence in LMs as investigative tools, we might start to use their behavior with specific languages as evidence for the possible/impossible status for those languages. In this way, LMs could also help us expand our evidence base.

Importantly, on this approach, the LM acts as a tool for informing theories of language and cognition via linking hypotheses, and its value as a tool is measured by the power of those linking hypotheses. Various other tools could, in principle, play the role of the LM. For example, simple classifiers defined over the languages from Phase 1 could provide insights. LMs are privileged only insofar as they are the most successful language technologies ever created, and they invite many architectural modifications that could support viable linking hypotheses.

It seems likely to us that using LMs in this context would reopen some of the usual terms of the debate. For example, LMs do not inherently define a binary notion of "learning a language." To address this, we could (1) posit such a distinction for them, (2) change them in fundamental ways, or (3) change our conception of what it means to learn a language, to allow for more gradient notions of learning. Our own view is that human language learning is itself gradient and fluid, so we would likely opt for (3), but other researchers might respond differently. This is a question of how best to use LMs as tools here, and we think raising such questions is itself productive.

We are energized by the above project, and we already see evidence that it is yielding new insights. For example, as noted above, prior work suggests that the very general learning mechanisms used by today's LMs suffice to create some observed linguistic locality effects. This is in itself illuminating about the factors that can lead to these locality effects. What sort of modifications to those mechanisms would increase this alignment? In addition, Hunter (2025) argues for the importance of linguistic constituent structure in any discussion of locality effects, and he describes new candidate possible/impossible language pairs designed to engage with this issue. What class of LMs is able to capture this asymmetry, and what might such LMs tell us about language and cognition?

Financial support. Julie Kallini is supported by a National Science Foundation Graduate Research Fellowship under grant number DGE-2146755.

Competing interests. The authors declare none.

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

References

Hunter, T. (2025). Kallini et al. (2024) do not compare impossible languages with constituency-based ones. *Computational Linguistics*, *51*(2), 641–650. https://doi.org/10.1162/coli_a_00554

Kallini, J., Papadimitriou, I., Futrell, R., Mahowald, K., & Potts, C. (2024). Mission: Impossible language models. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers) (pp. 14691–14714). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.787

McIntosh, L., Maheswaranathan, N., Nayebi, A., Ganguli, S., & Baccus, S. (2016). Deep learning models of the retinal response to natural scenes. In D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 29* (pp. 1369–1377). Curran Associates. https://proceedings.neurips.cc/paper_files/paper/2016/hash/a1d33d0dfec820b41b54430b50e96b5c-Abstract.html

Mitchell, J., & Bowers, J. (2020). Priorless recurrent networks learn curiously. In D. Scott, N. Bel, & C. Zong (Eds.), *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 5147–5158). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.451

Moro, A. (2008). *The boundaries of Babel: The brain and the enigma of impossible languages*. MIT Press. https://doi.org/10.7551/mitpress/9780262134989.001.0001

Moro, A., Greco, M., & Cappa, S. F. (2023). Large languages, impossible languages and human brains. *Cortex*, *167*, 82–85. https://doi.org/10.1016/j.cortex.2023.07.003

Xu, T., Kuribayashi, T., Oseki, Y., Cotterell, R., & Warstadt, A. (2026). Can language models learn typologically implausible languages? *Transactions of the Association for Computational Linguistics*, *14*, 588–611. https://doi.org/10.1162/TACL.a.640

Yang, X., Aoyama, T., Yao, Y., & Wilcox, E. (2025). Anything goes? A cross-linguistic study of (im)possible language learning in LMs. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 26058–26077). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.1264

Ziv, I., Lan, N., Chemla, E., & Katzir, R. (2026). Large language models as proxies for theories of human linguistic cognition. *Journal of Language Modelling*, *14*(2). https://doi.org/10.15398/jlm.457

Rich data drive generalization: Lessons from machine learning for linguistics and cognitive science

Andrew Kyle Lampinen*

Google DeepMind, USA
lampinen@google.com

*Corresponding author.

doi:10.1017/S0140525X26104609, e212

Abstract

The diversity of variation captured in data can strongly affect the generalization of a learning system – even when that variation occurs along axes orthogonal to the generalization in question. Thus, I argue that data richness both distinguishes current language models from prior linguistic models and may still underlie their remaining linguistic data inefficiency.

Language model training data are distinct not only in their quantity. In comparison to classic linguistic and cognitive models, language model training data are fundamentally richer, more diverse, and much noisier. Yet, they are still missing much of the diversity of human learning experiences. While Futrell & Mahowald discuss the role of data *quantity* in model training, they do not focus on the important role of data *richness*: that the diversity of *variation* captured in data can strongly affect the generalization of a learning system (cf. Raviv, Lupyan, & Green, 2022). This principle that learning diversity drives generalization has been frequently observed in machine learning, within and beyond the language domain (e.g., del Rio et al., 2025; Goyal et al., 2022; Yu, Khadivi, & Xu, 2022).

One illustrative example of this can be found in some of the target authors' own work, though it is not emphasized as such in the target article: Misra and Mahowald (2024) show that under the same train-test *syntactic* generalization split, variability in the *semantic fillers* observed in training data affects the degree of *syntactic* generalization. That is, variation along non-syntactic axes can enhance generalization to novel syntactic structures. (Of course, seeing greater variability in syntactic structures will also typically improve generalization; e.g., Ahuja et al., 2025; Qin, Saphra, & Alvarez-Melis, 2025).

Our prior work (Hill et al., 2020) on compositional generalization of language-learning agents shows a more dramatic effect: agents trained to perform classification through interaction in 3D environments show strong *compositional* language generalization, while agents trained to perform the same tasks on 2D images from the same environments generalize much more poorly. Thus, when comparing two language-learning systems – both of which are grounded, in the sense that their training tasks involve producing responses that depend on linking language to visual inputs and action outputs – the richness of the environmental grounding can significantly alter *linguistic* generalization. At a more basic level, various works have found that providing some type of semantic grounding can enhance syntactic generalization compared to systems that lack such grounding (e.g., Yedetore & Kim, 2024).

These works showing the contribution of data diversity to generalization in controlled experimental settings complement the many works showing its importance in practical language model training. For example, targeting sampling of training data for diversity can improve out-of-distribution perplexity (Yu et al., 2022). Conversely, language model training data collected from the internet often lacks diversity; for example, Lee et al. (2022) highlight a single sentence that appears over 60,000 times in a standard training corpus. What might naively seem to be 3.5 million tokens of unique training data (almost a year of a child's experience) turns out to be only 61 words of unique content – an important caveat to remember when seeing naive numeric comparisons of learning efficiency. While this is a particularly egregious example, the problem of partially duplicated training data is prevalent, and Lee et al. (2022) show that reducing this duplication can substantially improve performance as well as training efficiency. This work provides another example of how learning diversity is important in practice as well as in theory: repeating data can actually harm performance.

These works illustrate how – both in controlled settings and in practice when training language models – richer variation in learning experiences can fundamentally change the resulting patterns of generalization. Importantly, this is true both for variation along the axes on which generalization will be assessed (e.g., the diversity of syntactic structures seen), but also along axes of variation that seem orthogonal to the syntactic or compositional language generalization in question (e.g., the visual richness of multimodal inputs).

These findings have important theoretical implications for the linguistic and cognitive issues at play in the target article. In particular, they suggest that at least some of the benefit of data *quantity* for model performance may really be a benefit of data *diversity*, which only happens to be achieved through scaling the datasets. In itself, I believe this lesson is important for linguistics and cognitive science: how many famous negative results about neural networks (or other models) failing to generalize were wild extrapolations from training models on toy data that lacked all the richness of human experiences? We need better theories of how diversity contributes to generalization, and we should discount models that neglect the diversity of human learning experiences without a compelling reason.

Yet if language models benefit from data diversity, why do they still need so much more data than humans? Of course, many factors may contribute; choices like data-deduplication strategies, tokenization, and optimizers can all drastically change learning efficiency; we probably have not saturated improvements along any of these dimensions. However, I suggest that differences in the richness of the input humans experience may play a crucial role. Human linguistic inputs begin as auditory signals rather than discrete tokens; these signals carry much more information to predict. Furthermore, human multimodal grounding is much richer. Indeed, while many models are now trained in part with multimodal inputs, for various reasons, language still tends to dominate (e.g., Sim et al., 2025) – whereas for humans, language develops more slowly than other systems for interaction. Indeed, there have been many arguments that features like prosody (e.g., Gervain, Christophe, & Mazuka, 2020), gesture (e.g., Iverson & Goldin-Meadow, 2005), and joint attention (Tomasello & Farrar, 1986) play an important role in human language development. Current language model training paradigms lack these, and many other rich features of human learning. However, simulating richer learning experiences is increasingly achievable. Thus, I believe that the modern paradigm of data-driven machine learning opens up many promising directions for exploring the role of linguistic and nonlinguistic data features in language acquisition (cf. Carvalho & Lampinen, 2025).

Acknowledgments. I thank Felix Hill for his mentorship and many conversations on these issues, Michael Mozer for comments and suggestions, and Michael Terry for support.

Financial support. None.

Competing interests. Andrew Lampinen is employed by Google DeepMind.

References

Ahuja, K., Balachandran, V., Panwar, M., He, T., Smith, N. A., Goyal, N., & Tsvetkov, Y. (2025). Learning syntax without planting trees: Understanding hierarchical generalization in transformers. *Transactions of the Association for Computational Linguistics*, *13*, 121–141.

Carvalho, W., & Lampinen, A. (2025). Naturalistic Computational Cognitive Science: Towards generalizable models and theories that capture the full range of natural behavior. *arXiv preprint arXiv:2502.20349*.

del Rio, F., Raymond-Sáez, A., Florea, D., Icarte, R. T., Hurtado, J., Calderon, C. B., & Soto, A. (2025). Data distributional properties as inductive bias for systematic generalization. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (pp. 25590–25601).

Gervain, J., Christophe, A., & Mazuka, R. (2020). Prosodic bootstrapping. *The Oxford Handbook of Language Prosody*, 563–573.

Goyal, P., Duval, Q., Seessel, I., Caron, M., Misra, I., Sagun, L., Joulin, A., & Bojanowski, P. (2022). Vision models are more robust and fair when pretrained on uncurated images without supervision. arXiv preprint arXiv:2202.08360.

Hill, F., Lampinen, A., Schneider, R., Clark, S., Botvinick, M., McClelland, J. L., & Santoro, A. (2020) Environmental drivers of systematicity and generalization in a situated agent. In *Proceedings of the International Conference on Learning Representations*.

Iverson, J. M., & Goldin-Meadow, S. (2005). Gesture paves the way for language development. *Psychological Science*, *16*(5), 367–371.

Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., & Carlini, N. (2022, May). Deduplicating training data makes language models better. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 8424–8445).

Misra, K., & Mahowald, K. (2024). Language models learn rare phenomena from less rare phenomena: The case of the missing AANNs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 913–929). Association for Computational Linguistics.

Qin, T., Saphra, N., & Alvarez-Melis, D. (2025). Sometimes I am a tree: Data drives unstable hierarchical generalization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing 2025 Nov* (pp. 11733–11751).

Raviv, L., Lupyan, G., & Green, S. C. (2022). How variability shapes learning and generalization. *Trends in Cognitive Sciences*, *26*(6), 462–483.

Sim, M. Y., Zhang, W. E., Dai, X., & Fang, B. (2025). Can vlms actually see and read? A survey on modality collapse in vision-language models. In *Findings of the association for computational linguistics: ACL 2025* (pp. 24452–24470). Association for Computational Linguistics.

Tomasello, M., & Farrar, M. J. (1986). Joint attention and early language. *Child development*, *57*(6), 1454–1463.

Yedetore, A., & Kim, N. (2024). Semantic training signals promote hierarchical syntactic generalization in transformers. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 4059–4073).

Yu, Y., Khadivi, S., & Xu, J. (2022). Can data diversity enhance learning generalization?. In *Proceedings of the 29th International Conference on Computational Linguistics* (pp. 4933–4945).

Large language models have learned to use language

Gary Lupyan*

Department of Psychology, University of Wisconsin-Madison, Madison, USA
lupyan@wisc.edu

doi:10.1017/S0140525X26104804, e213

Abstract

Acknowledging that large language models have learned to use language can open doors to breakthrough language science. Achieving these breakthroughs may require abandoning some long-held ideas about how language knowledge is evaluated and reckoning with the difficult fact that we have entered a post-Turing test era.

It is instructive to compare Futrell and Mahowald's well-reasoned and empirically grounded target article with counterpoints from skeptics demanding that large language models (LLMs) conform to theories of generative linguistics to be judged as good models of language (Bolhuis et al., 2024; Murphy et al., 2025), or who insist that LLMs are "stochastic parrots" that "only manipulate the form of language, with neither understanding nor communicative intent" (Bender & Hanna, 2025). Gauging the usefulness of LLMs as models of language based on how well they adhere to a particular theory of language is not only self-serving and unscientific but also risks further isolating many linguists from the broader language sciences. And reducing LLMs to "stochastic parrots" is an insult to people's intelligence. Are we to believe that the 800M weekly users of ChatGPT are credulous dolts unable to distinguish competent language use from mere parroting? Or is it more likely that a decades-long scientific program of learning language through prediction (Elman, 1990) was on to something – that learning language from data is not only possible but actually works?

By "actually works", I mean that a general-purpose neural network endowed with a simple learning rule to lower prediction error, an architectural capacity to track long-distance associations, and a corpus of language use can learn language. Some continue to deny the premise, arguing that the failure of LLMs to "Draw [a] tree structure, in line with Minimalist syntax, for the sentence 'The doctor the nurse the hospital had hired met John'" means that LLMs lack "fundamental principles of linguistic structure" (Murphy et al., 2025). This argument is absurd. My 5-year-old is a highly competent language user who cannot draw a parse tree or tell you what makes a sentence ungrammatical. What counts is use! If a language model can use language proficiently enough to pass the Turing test (Jones & Bergen, 2025) while failing to conform to minimalist syntax, so much the worse for minimalist syntax (Hu et al., 2024).

The ability to judge sentences like "Colorless green ideas sleep furiously" to be grammatical despite being nonsensical has played an outsized role in generative linguistics. These metalinguistic judgments are often used to test competing theories (Schütze, 2016) and have often been used to evaluate LLMs (e.g., Dentella, Günther, & Leivada, 2023). That children engage in sophisticated language use while lacking the ability to make such explicit judgments is *prima facie* evidence that being a competent language user does not require being able to dissociate conceptual content from structural configurations. It is also true, however, that many adults without linguistic training can do this. But so can LLMs. As we previously showed, ChatGPT-4 is better than people at judging the normative grammaticality of sentences (Hu et al., 2024). In recent informal tests, we also found that GPT-5 can *generate* meaningless grammatical sentences as well as adults. Figure 1 shows meaningfulness ratings of sentences generated by people and GPT-5 ($t < 1$), as well as people's attempts to guess whether each sentence was generated by another person or an LLM ($t < 1$).

An intriguing question is when the ability to distinguish meaningfulness from grammaticality emerges in LLMs and whether this ability has any causal bearing on proficient language use. Although grammaticality can be read out of base models (LLMs trained only through self-supervision) using surprisal, base models appear incapable of making explicit grammaticality judgments (Hu et al., 2024). In our preliminary tests, we find that base models can generate implausible sentences (e.g., A cat barks at the dog) but struggle to generate ungrammatical or truly meaningless ones. The ability to explicitly uncouple meaningfulness from grammaticality appears to emerge with fine-tuning on even small datasets (e.g., Databricks-dolly-15k, 2023) which do not contain any examples of grammaticality judgments. What makes this intriguing is that despite their enormous latent knowledge of both language and the world, base models make lousy conversationalists (the public couldn't care less about GPT-3 but was blown away by its fine-tuned version released as ChatGPT). Understanding whether the ability to make explicit grammatical judgments is causally related to being able to *use* language would be a fascinating development.

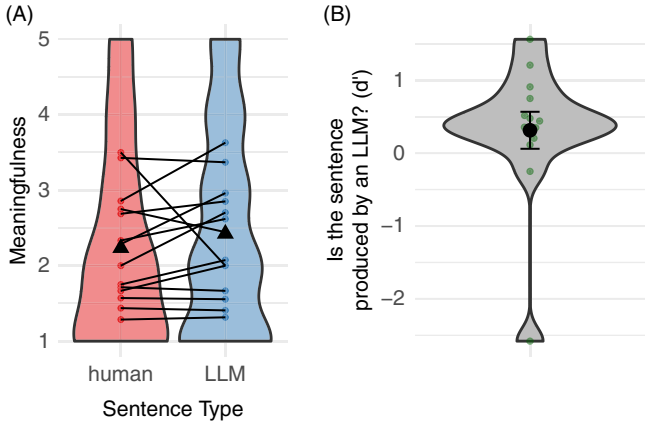


Figure 1 (Lupyan). A. Results from an informal experiment in which participants were tasked with generating maximally meaningless grammatical sentences. Each participant then judged the meaningfulness of the sentences generated by other participants and ChatGPT-5 and whether each sentence was generated by a person or an LLM. (A) Mean meaningfulness ratings with lines joining each participant's mean ratings. (B) Accuracy in judging whether each sentence was generated by a person or LLM. An example of similarly meaningless LLM and human sentences: "Let velvet measure tomorrow with raisins" and "Burpy trees monkey in mountain poodle".

It seems that over the decades, some linguists have lost the plot (see Knight, 2016 for an engaging intellectual history). The goal of a language user is not to generate parse trees that conform to minimalist syntax. Rather, it is to *do* things with language. This includes inferring people's mental states from what they say, communicating goals to enable collaboration, and convincing them to act in your interests. To the extent that such language use *requires* learning specific aspects of language structure, we should find these by appropriately interrogating LLMs. For example, it seems clear that representing the hierarchical structure – in some form – is critical to parsing language, and converging evidence appears to show that in learning language, LLMs learn hierarchical structure (e.g., Goldstein et al., 2025; Liu, Xiang, & Ding, 2026). LLMs provide an unprecedented opportunity to understand that the aspects of linguistic structure are critical for language use, even when learned by a system with an architecture radically different from ours.

In his 1990 short story *Bears Discover Fire*, Terry Bisson imagines a world just like ours, but in which, inexplicably, bears have discovered fire: "They don't hibernate anymore ... They make a fire and keep it going all winter" (Bisson, 1990). One reaction to such a development would be to insist that bears are just not the kind of animal that can control fire; that it's a trick; that their fire must not be *real* fire. Another would be to wonder if perhaps we were mistaken in our understanding of what bears are capable of, or in our understanding of what it takes to control fire. We have now entered a post-Turing test era where the torch of language is no longer unique to our species. What we make of this new reality is up to us.

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

Acknowledgments. The author would like to thank Zach Studdiford for his help on comparing Gemma base and instruction-tuned language models.

Financial support. This work was partially supported by NSF-PAC 2020969 to G.L.

Competing interests. None.

References

Bender, E. M., & Hanna, A. (2025). *The AI con: How to fight big tech's hype and create the future we want*. Harper.

Bisson, T. (1990, 2014). Bears discover fire. *Lightspeed Magazine*, *44*. https://www.lightspeedmagazine.com/fiction/bears-discover-fire/

Bolhuis, J. J., Crain, S., Fong, S., & Moro, A. (2024). Three reasons why AI doesn't model human language. *Nature*, *627*(8004), 489–489. https://doi.org/10.1038/d41586-024-00824-z

Databricks/databricks-dolly-15k · Datasets at Hugging Face. (2023, October 1). https://huggingface.co/datasets/databricks/databricks-dolly-15k

Dentella, V., Günther, F., & Leivada, E. (2023). Systematic testing of three Language Models reveals low language accuracy, absence of response stability, and a yes-response bias. *Proceedings of the National Academy of Sciences*, *120*(51), e2309583120. https://doi.org/10.1073/pnas.2309583120

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

Goldstein, A., Ham, E., Schain, M., Nastase, S. A., Aubrey, B., Zada, Z., Grinstein-Dabush, A., Gazula, H., Feder, A., Doyle, W., Devore, S., Dugan, P., Friedman, D., Brenner, M., Hassidim, A., Matias, Y., Devinsky, O., Siegelman, N., Flinker, A., Levy, O., Reichart, R., & Hasson, U. (2025). Temporal structure of natural language processing in the human brain corresponds to layered hierarchy of large language models. *Nature Communications*, *16*(1), 10529. https://doi.org/10.1038/s41467-025-65518-0

Hu, J., Mahowald, K., Lupyan, G., Ivanova, A., & Levy, R. (2024). Language models align with human judgments on key grammatical constructions. *Proceedings of the National Academy of Sciences*, *121*(36), e2400917121. https://doi.org/10.1073/pnas.2400917121

Jones, C. R., & Bergen, B. K. (2025). *Large Language Models Pass the Turing Test* (No. arXiv:2503.23674). arXiv. https://doi.org/10.48550/arXiv.2503.23674

Knight, C. (2016). *Decoding chomsky: Science and revolutionary politics*. Yale University Press.

Liu, W., Xiang, M., & Ding, N. (2026). Active use of latent tree-structured sentence representation in humans and large language models. *Nature Human Behaviour*, *10*(2), 303–316. https://doi.org/10.1038/s41562-025-02297-0

Murphy, E., Leivada, E., Dentella, V., Montero, R., Günther, F., & Marcus, G. (2025). Fundamental principles of linguistic structure are not represented by ChatGPT. *Biolinguistics*, *19*, 1–55. https://doi.org/10.5964/bioling.19021

Schütze, C. T. (2016). *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Language Science Press. https://www.oapen.org/search?identifier=603356

Not-so-strange love: Language models and generative linguistic theories are more compatible than they appear

R. Thomas McCoy*

Yale University, USA
tom.mccoy@yale.edu

*Corresponding author.

doi:10.1017/S0140525X26104750, e214

Abstract

Futrell and Mahowald frame the success of neural language models (LMs) as supporting gradient, usage-based linguistic theories. I argue that LMs can also instantiate theories based on formal structures – the types of theories seen in the generative tradition. This argument expands the space of theories that can be tested with LMs, potentially enabling reconciliations between usage-based and generative accounts.

What role does formal structure play in language acquisition? This question is at the heart of longstanding debates in linguistics. Generative linguists argue that structure guides learning through innate predispositions for certain formal properties (Chomsky, 1993). Usage-based linguists instead argue that the central factor is statistics of language use, with formal structure emerging instead of being innate (Bybee & Hopper, 2001). Futrell and Mahowald align language models (LMs) with the usage-based tradition, framing them as "a proof of concept for gradient, usage-based theories of language." Here, I argue that while LMs can serve this purpose, they can also instantiate theories based on formal structures. Therefore, they are compatible with generative theories as well as usage-based ones, and their success does not uniquely vindicate one or the other.

As Futrell and Mahowald discuss, standard LMs possess many properties that align with usage-based theories, such as their statistical nature and lack of built-in structural constraints. Thus, LMs indeed provide partial support for usage-based ideas. However, this support is only partial due to an important limitation of standard LMs: as Futrell and Mahowald note, LMs are typically trained on far more linguistic input than children receive. When they instead receive more realistic amounts of data, their linguistic abilities are less impressive (Yedetore et al., 2023). Such limitations could potentially be overcome by predisposing LMs toward formal linguistic structures, since a major motivation for centering formal structure is to explain how children acquire language from so little data.

In fact, there is an approach that can distill a formal-structure-based linguistic theory into a neural network, enabling LMs to serve as empirical testing grounds for generative theories as well as usage-based ones (McCoy & Griffiths, 2025). This approach uses a technique called meta-learning, in which the network is shown many languages sampled from a space of possible languages. Through this process, the network learns which sorts of languages are likely and unlikely, endowing it with a version of Universal Grammar. This information provides helpful guidance for language learning, enabling the network to learn new languages more readily.

For example, in McCoy et al., (2020), we used this approach to distill an Optimality Theory account of syllable structure (Prince & Smolensky, 1993/2004) into a neural network, giving it learning biases that instantiated this generative theory. The resulting network could learn syllable structure patterns from just 100 examples, compared to the 20,000 examples needed by standard networks. Further, the network generalized well to novel types of examples, such as words longer than any it had seen. Thus, after having a generative theory distilled into it, the network displayed precisely the strengths that structure-centric learning theories are intended to capture: rapid learning and effective extrapolation – areas where standard neural networks struggle. In addition to this

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

phonological case study, we have applied the same technique to syntax, showing that learning biases for certain formal language mechanisms can be distilled into LMs (McCoy & Griffiths, 2025). These experiments show that we can create neural networks with a decidedly generative flavor, in that predispositions for specific formal structures drive their learning.

The fact that LMs can instantiate generative theories raises the possibility that existing LMs might already implicitly realize aspects of these theories, further complicating the question of whether existing LM successes are driven by usage-based or structural properties. Structural biases could arise from architectural factors (Smolensky et al., 2022) or from highly structured training data such as computer code (Kim, Schuster, & Toshniwal, 2024), which might contribute formal structural biases through a process akin to meta-learning. Whether existing LMs indeed have formal structural influences, and whether the relevant structures align with existing generative theories, are important questions for future research.

Because LMs can instantiate usage-based theories or generative ones, they create opportunities to evaluate both types of theories. Further, the gradient nature of LMs means that they are not restricted to incorporating only one type of theory but can instead combine usage-based and generative properties. Therefore, they create a pathway for exploring how we can combine the complementary strengths of these two paradigms (Boleda, 2025).

As one example of synthesizing the two schools of thought, consider the aforementioned syllable structure experiment, which instantiated a generative theory inside a neural network. The instantiation's neural nature meant that distinctions that were absolute in the generative theory became graded in the neural instantiation. For instance, under the generative theory in question, certain types of languages are impossible, but our neural instantiation of the theory could still learn such languages; "impossible" languages required far more training data than "possible" ones, but they could eventually be learned. Thus, the notion of impossibility in the generative theory became improbability inside its neural instantiation. There are at least two ways that this inconsistency can be viewed as combining generative ideas and gradient, usage-based ones. First, the network can be viewed as combining the causal influence of formal structure from the generative tradition with the gradience typically associated with usage-based theories. Under this framing, the network is neither fully generative nor fully aligned with the usage-based tradition but rather combines elements of both. Alternatively, the network could be viewed as affording two levels of analysis in the sense of Marr (1982). On one level – the level defined by the neural network, corresponding to Marr's algorithmic level – the system is gradient, while at a more abstract, idealized level – the level defined by the generative theory that the network was optimized to realize, corresponding to Marr's computational level – the system is discrete. This framing illustrates how a single system can be understood in different ways at different levels (Smolensky & Legendre, 2006), as another way in which seemingly incompatible theoretical traditions can be reconciled.

There is precedent for generative linguistics being deeply shaped by neural networks. Optimality Theory, an influential generative formalism, was motivated by neural networks, demonstrating that neural networks and generative theories can interact productively. Recent progress in neural LMs creates further opportunities for advancing generative linguistics and for exploring syntheses between generative and usage-based theories.

Acknowledgements. I am grateful to Robert Frank and Kate McCurdy for their helpful discussion. Any errors are my own.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing interests. None.

References

Boleda, G. (2025). LLMs as a synthesis between symbolic and distributed approaches to language. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 9365-9379.

Bybee, J. L. & Hopper, P. (2001). *Frequency and the Emergence of Linguistic Structure*. Typological Studies in Language 45. John Benjamins.

Chomsky, N. (1993). A minimalist program for linguistic theory. In K. Hale, & S.J. Keyser (Eds.), *The view from building 20* (pp. 1–52). MIT Press.

Kim, N., Schuster, S., & Toshniwal, S. (2024). Code pretraining improves entity tracking abilities of language models. *arXiv preprint arXiv:2405.21068*.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W.H. Freeman.

McCoy, R. T., Grant, E., Smolensky, P., Griffiths, T. L., & Linzen, T. (2020). Universal linguistic inductive biases via meta-learning. *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*, 737–743.

McCoy, R. T., & Griffiths, T. L. (2025). Modeling rapid language learning by distilling Bayesian priors into artificial neural networks. *Nature Communications*, *16*(1), 4676.

Prince, A., & Smolensky, P. (1993/2004). *Optimality theory: Constraint interaction in generative grammar*. Wiley.

Smolensky, P., & Legendre, G. (2006). *The harmonic mind: From neural computation to optimality-theoretic grammar*. MIT.

Smolensky, P., McCoy, R., Fernandez, R., Goldrick, M., & Gao, J. (2022). Neurocompositional computing: From the central paradox of cognition to a new generation of AI systems. *AI Magazine*, *43*(3), 308–322.

Yedetore, A., Linzen, T., Frank, R., & McCoy, R. T. (2023). How poor is the stimulus? Evaluating hierarchical generalization in neural networks trained on child-directed speech. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9370–9393.

Beyond the data gap: Children create languages, violate their input statistics, and exhibit critical periods

Annika McDermott-Hinman[a] and Roman Feiman[a,b]*

[a]Cognitive and Psychological Sciences, Brown University, Providence, USA and [b]Linguistics, Brown University, Providence, USA
annikamh406@gmail.com
roman_feiman@brown.edu

*Corresponding author.

doi:10.1017/S0140525X26104671, e215

Abstract

Futrell and Mahowald argue language model (LM) success suggests humans may learn language entirely through domain-general statistical mechanisms. However, children differ crucially

from LMs in their ability to surpass their input, their language learning trajectory, and the presence of a critical period. Until LMs account for these phenomena, it remains possible that human language acquisition is supported by innate, language-specific learning mechanisms.

Futrell and Mahowald argue that because language models (LMs) can learn linguistic structure from text input without explicitly encoded, language-specific inductive biases, we need not suppose children have such biases either. This argument requires that LMs learn linguistic structure in the same way as children; the only difference Futrell and Mahowald consider is how much data each needs to reach mature competence (see also Frank, 2023). Yet LMs and children also systematically diverge in *how* they acquire language, and even in what it is they acquire. If LMs cannot eventually model these aspects of acquisition, language-specific learning mechanisms may still be needed to explain them.

Language emergence. When children receive sparse linguistic input, they spontaneously generate linguistic structure not present in that input. Deaf children not exposed to sign language develop homesign systems whose structure and expressiveness outstrip the gestures of their hearing families. Homesign systems morphologically or syntactically distinguish parts of speech (Abner et al., 2019; Goldin-Meadow et al., 1994) and thematic roles (Goldin-Meadow & Mylander, 1998) and structurally mark grammatical subjects (Coppola & Newport, 2005) – distinctions absent or much reduced in homesigners' hearing families' gestures.

Further, when children's input is limited to a rudimentary linguistic system like homesign, they quickly expand and regularize it. Famously, the conventionalized sign system generated de novo by children attending a Nicaraguan school for the deaf was elaborated by successive student cohorts, each imposing additional structure absent from their input. For instance, the second cohort, but not the first, uses spatial modulation to convey shared reference (Senghas & Coppola, 2001; Senghas, 2003). Over a few cohorts, Nicaraguan Sign Language has come to exhibit discrete, compositional morphology (Senghas, Kita, & Özyürek, 2004) and recursion (Kocab et al., 2023). Much of the structure of emerging languages originates from the child, not their input.

In contrast, LMs learn only to recapitulate the statistical properties of their training data. To account for language emergence – that is, the addition of linguistic structure that is absent from the input – LMs would need to *violate* those statistical properties. If that is possible, LMs would then need to more specifically enrich the structure of a language in the ways children do when placed under similar communicative pressures. Although simple neural agents can create rudimentary compositional systems in cooperative reference games (Boldt & Mortensen, 2024; Lazaridou & Baroni, 2020; Steinert-Threlkeld, 2020), it is uncertain whether and under which conditions LMs could create systems with the rich structural complexity that emerges in the ad hoc communication system of even a single child.

Learning trajectories. Children's word-learning trajectories exhibit characteristic patterns distinct from the frequency of words in their input. Concrete nouns are overrepresented at first, then verbs and adjectives, and only later do closed-class function words like "of" and "the" follow – though these are the most frequent words in the input (Frank et al., 2016). Further, even the most basic syntactic productivity (e.g., combining two words) emerges only after a period when children communicate meaningfully using early words in isolation (Fenson et al., 2007). LMs show the opposite pattern: they produce correct complex multi-word syntax early in training, but semantically coherent productions only much later (Müller-Eberstein et al., 2023; Xia et al., 2023). And unlike children, LMs' early productions reflect token input frequencies. "The," periods, and commas dominate (Chang, Tu, & Bergen, 2024).

These learning trajectories reflect fundamental differences in how children and LMs use language. LMs learn patterns of co-occurrence between constituents of their text-based input. Children must additionally learn how words *refer* to the world. This is why children first learn nouns for concrete objects: those are the easiest to identify as referents in their environment (Gillette et al., 1999; Snedeker, Geren, & Shafto, 2007). This gap might be bridged by multimodal LMs that learn both word-word and word-percept relations. Yet early attempts show that such models have severe limitations in even the simplest referential abilities, such as generalizing a learned label-object pairing to new objects of the same shape (Vong et al., 2024). While more training data may help, these limitations may reflect deeper differences in how LMs and children acquire the concepts underlying word meanings. Just as evidence from language emergence shows that children add morphosyntactic structure absent from their linguistic input, scholars from Kant to Spelke to Gleitman have adduced evidence that children's preverbal conceptual and perceptual capacities structure thought beyond the combined associations of words with other words and with images. To the extent that learning language is the process of learning the expression of thought, LMs would need to model humans' capacities for thought to fully model their language.

Critical periods. While general statistical learning abilities improve throughout development (Shufaniya & Arnon, 2018), children's ability to learn the grammar of their language declines around puberty (Hartshorne, Tenenbaum, & Pinker, 2018; Johnson & Newport, 1989; Lenneberg, 1967), and the ability to learn native phonology declines in infancy (see Werker, 2024, for review). This dissociation implies that learning language does not rely solely on general statistical mechanisms.

LMs might account for critical periods by demonstrating that declines in language learning ability can arise from language-specific experience that would not impede general statistical learning. However, recent work has found the opposite: learning a first language *accelerates* the LM's rate of subsequently learning a second (Oba et al., 2023). In fact, the way that a critical period *can* be induced in LMs is by imposing a regularizer partway through training – a process more analogous to a maturational decrease in language-specific plasticity than to any effect of experience (Constantinescu et al., 2025). By contrast, neural networks in other domains do show spontaneous critical periods. For instance, deep multisensor perceptual networks show a critical period during which improperly correlated data can permanently impair the model's learning (Kleinman, Achille, & Soatto, 2023). Thus, while neural networks may be able to account for critical periods in some areas of development based on experience alone, the way LMs have been able to account for them in language specifically favors a maturational, language-specific explanation.

Acknowledgments. Our thanks to Ann Senghas for insightful comments and suggestions on an earlier draft and to Marie Coppola for helpful discussion.

Financial support. This work was supported in part by NSF CAREER award #2442374 to R.F.

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

Competing interests. None.

References

Abner, N., Flaherty, M., Stangl, K., Coppola, M., Brentari, D., & Goldin-Meadow, S. (2019). The noun-verb distinction in established and emergent sign systems. *Language*, *95*(2), 230–267. https://doi.org/10.1353/lan.2019.0030

Boldt, B., & Mortensen, D. (2024). *A Review of the Applications of Deep Learning-Based Emergent Communication* (No. arXiv:2407.03302). arXiv. https://doi.org/10.48550/arXiv.2407.03302

Chang, T. A., Tu, Z., & Bergen, B. K. (2024). Characterizing learning curves during language model pre-training: Learning, forgetting, and stability. *Transactions of the Association for Computational Linguistics*, *12*, 1346–1362. https://doi.org/10.1162/tacl_a_00708

Constantinescu, I., Pimentel, T., Cotterell, R., & Warstadt, A. (2025). Investigating critical period effects in language acquisition through neural language models. *Transactions of the Association for Computational Linguistics*, *13*, 96–120. https://doi.org/10.1162/tacl_a_00725

Coppola, M., & Newport, E. L. (2005). Grammatical Subjects in home sign: Abstract linguistic structure in adult primary gesture systems without linguistic input. *Proceedings of the National Academy of Sciences*, *102*(52), 19249–19253. https://doi.org/10.1073/pnas.0509306102

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, J. S., & Bates, E. (2007). *MacArthur-bates communicative development inventories: User's guide and technical manual* (2nd ed.) Brookes Publishing Co.

Frank, M. C. (2023). Bridging the data gap between children and large language models. *Trends in Cognitive Sciences*, *27*(11), 990–992. https://doi.org/10.1016/j.tics.2023.08.007

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2016). Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language*, *44*(3), 677–694. https://doi.org/10.1017/S0305000916000209

Gillette, J., Gleitman, H., Gleitman, L., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition*, *73*, 135–176.

Goldin-Meadow, S., Butcher, C., Mylander, C., & Dodge, M. (1994). Nouns and verbs in a self-styled gesture system: What's in a name? *Cognitive Psychology*, *27*, 259–319.

Goldin-Meadow, S., & Mylander, C. (1998). Spontaneous sign systems created by deaf children in two cultures. *Nature*, *391*(6664), 279–281. https://doi.org/10.1038/34646

Hartshorne, J. K., Tenenbaum, J. B., & Pinker, S. (2018). A critical period for second language acquisition: Evidence from 2/3 million English speakers. *Cognition*, *177*, 263–277. https://doi.org/10.1016/j.cognition.2018.04.007

Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, *21*(1), 60–99. https://doi.org/10.1016/0010-0285(89)90003-0

Kleinman, M., Achille, A., & Soatto, S. (2023). Critical learning periods for multisensory integration in deep networks. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 24296–24305. https://doi.org/10.1109/CVPR52729.2023.02327

Kocab, A., Senghas, A., Coppola, M., & Snedeker, J. (2023). Potentially recursive structures emerge quickly when a new language community forms. *Cognition*, *232*, 105261. https://doi.org/10.1016/j.cognition.2022.105261

Lazaridou, A., & Baroni, M. (2020). *Emergent Multi-Agent Communication in the Deep Learning Era* (No. arXiv:2006.02419). arXiv. https://doi.org/10.48550/arXiv.2006.02419

Lenneberg, E. H. (1967). *Biological foundations of language*. John Wiley & Sons.

Müller-Eberstein, M., Van Der Goot, R., Plank, B., & Titov, I. (2023). Subspace chronicles: How linguistic information emerges, shifts and interacts during language model training. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 13190–13208. https://doi.org/10.18653/v1/2023.findings-emnlp.879

Oba, M., Kuribayashi, T., Ouchi, H., & Watanabe, T. (2023). *Second Language Acquisition of Neural Language Models* (No. arXiv:2306.02920). arXiv. https://doi.org/10.48550/arXiv.2306.02920

Senghas, A. (2003). Intergenerational influence and ontogenetic development in the emergence of spatial grammar in Nicaraguan Sign Language. *Cognitive Development*, *18*(4), 511–531. https://doi.org/10.1016/j.cogdev.2003.09.006

Senghas, A., & Coppola, M. (2001). Children creating language: How Nicaraguan Sign Language acquired a spatial grammar. *Psychological Science*, *12*(4), 323–328. https://doi.org/10.1111/1467-9280.00359

Senghas, A., Kita, S., & Özyürek, A. (2004). Children creating core properties of language: Evidence from an emerging Sign Language in Nicaragua. *Science*, *305*(5691), 1779–1782. https://doi.org/10.1126/science.1100199

Shufaniya, A., & Arnon, I. (2018). statistical learning is not age-invariant during childhood: Performance improves with age across modality. *Cognitive Science*, *42*(8), 3100–3115. https://doi.org/10.1111/cogs.12692

Snedeker, J., Geren, J., & Shafto, C. L. (2007). Starting over: International adoption as a natural experiment in language development. *Psychological Science*, *18*(1), 79–87. https://doi.org/10.1111/j.1467-9280.2007.01852.x

Steinert-Threlkeld, S. (2020). Toward the emergence of nontrivial compositionality. *Philosophy of Science*, *87*(5), 897–909. https://doi.org/10.1086/710628

Vong, W. K., Wang, W., Orhan, A. E., & Lake, B. M. (2024). Grounded language acquisition through the eyes and ears of a single child. *Science*, *383*(6682), 504–511. https://doi.org/10.1126/science.adi1374

Werker, J. F. (2024). Phonetic perceptual reorganization across the first year of life: Looking back. *Infant Behavior and Development*, *75*, 101935. https://doi.org/10.1016/j.infbeh.2024.101935

Xia, M., Artetxe, M., Zhou, C., Lin, X. V., Pasunuru, R., Chen, D., Zettlemoyer, L., & Stoyanov, V. (2023). Training trajectories of language models across scales. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 13711–13738. https://doi.org/10.18653/v1/2023.acl-long.767

Keeping it real: Language models implement algorithms to solve linguistic tasks

Raphaël Millière*

Faculty of Philosophy, University of Oxford, UK
raphael.milliere@philosophy.ox.ac.uk

*Corresponding author.

doi:10.1017/S0140525X26104749, e216

Abstract

Futrell & Mahowald argue that language models (LMs) discover linguistic structure as "real patterns." I contend this framing underplays what mechanistic interpretability uncovers: LMs implement specific algorithms to solve linguistic tasks. Under the framework of causal abstraction, we can rigorously test whether LMs converge on algorithms posited by linguistic theory, which further supports the authors' conciliatory proposal.

Futrell & Mahowald (F&M) offer a compelling middle path in current debates about the relevance of language models (LMs) to linguistics. Against those who treat LMs as mere engineering tools with no bearing on linguistic theory, they argue that any system that solves the hard problem of language prediction and comprehension is likely to converge on a similar underlying structure. Against those who take the success of LMs to obviate linguistic

theory, they insist that linguistic structure remains indispensable as a "real pattern" in Dennett's sense: a higher-level abstraction that compresses, predicts, and supports counterfactual reasoning about both human and LM behaviour, even if it is not hard-wired. They further present mechanistic interpretability as a way to bridge such abstract structural descriptions and messy implementation details. I agree with this overall picture, but think it underplays what that bridge can carry. Not only do LMs discover real patterns in language, but they also implement specific algorithms for solving linguistic tasks. This algorithmic perspective strengthens F&M's conciliatory project.

On Dennett's account, a pattern is "real" if a description that uses it to explain a system's behaviour is significantly more compressive and predictively useful than any available lower-level description. F&M plausibly treat grammatical constructs such as "subject," "agreement," or "filler–gap dependency" as real patterns in this sense: they are higher-level descriptions that let us summarise and forecast linguistic behaviour far more efficiently than tracking neural activations would. But "real pattern" is an intentionally liberal notion: many stable statistical regularities will qualify. By contrast, linguistic theories are traditionally construed as hypotheses about algorithms. A grammar does not just encode which strings are licit; it specifies how structured representations are built and transformed. If we translate this into generic talk of "real patterns," we risk losing sight of what is distinctive about algorithmic explanation: it isolates the stepwise procedure that any adequate implementation must, in some sense, realise.

Mechanistic interpretability, as F&M emphasise, lets us uncover how LMs process information internally. They highlight work that identifies "circuits," or subgraphs of neural networks, that appear to be causally responsible for computing particular syntactic features such as number agreement. Describing how the components of these circuits interact, while omitting many lower-level implementation details, yields what we might call *mechanism sketches* for linguistic behaviour (Piccinini & Craver, 2011). Mechanism sketches of this kind are compressive and predictive. But algorithms are not mechanism sketches: they are formal, abstract, and multiply realisable procedures that stand in an implementation relation to particular circuits (Williams, 2025). To describe what a system does at the algorithmic level of analysis is to specify a particular sequence of representational states and computations through which it transforms inputs into outputs.

We can formalise algorithmic-level claims about LMs within the framework of causal abstraction (Geiger et al., 2025). An LM is represented as a low-level causal model, while an algorithm hypothesised to compute a given syntactic feature (e.g., subject–verb agreement) is represented as a high-level causal system whose nodes are linguistic variables (subject number, attractor number, verb form, etc.). We then ask: does there exist an alignment between them such that interventions on the LM's internal states have the same behavioural outcome as interventions on the variables of the algorithm? Concretely, this can be probed by interchange interventions. We feed the LM a base input and a counterfactual input that differ with respect to the target feature, then "patch" a candidate subspace of the LM's hidden activations in the base run with those from the counterfactual run. If that subspace really encodes the target feature (e.g., subject number), then patching it should change the model's output (e.g., predicted verb form) exactly as the algorithm predicts, across many sentences and many intervention sites. When a simple alignment of this kind achieves high interchange intervention accuracy, we have evidence that the LM implements the candidate algorithm.

Focusing on algorithms sharpens the relevance of LMs to linguistic theory, which offers hypotheses about algorithms that construct and constrain structured linguistic representations. Psycholinguistics proposes algorithms for how such representations are used in parsing, prediction, and retrieval. Interpretability under causal abstraction lets us ask whether an LM has in fact converged on those algorithms, or on some different procedure that nonetheless supports similar surface patterns. Either answer is informative: convergence would vindicate the algorithmic hypothesis as a robust solution to a hard problem, while divergence might lead us to reconsider what goes on in the human case. In addition, algorithmic-level explanations for the linguistic behaviour of LMs allow us to carve out a meaningful double dissociation between competence and performance in humans and machines alike (Millière & Rathkopf, 2026).

One might worry that talk of implementation is cheap: with a sufficiently baroque mapping between states, any network can be said to implement any algorithm. Causal abstraction can be supplemented to address this triviality concern in two ways (Geiger, Harding, & Icard, 2025). First, we can restrict the admissible mappings, for example, to reasonably simple linear transformations of activation space plus coarse-graining. Second, and more importantly, we can build in a demand for generalisation. A good candidate algorithm must predict the model's behaviour not only on inputs that resemble its training data but also under novel interventions and out-of-distribution cases.

In sum, F&M are right that linguistic structure should be treated as a real pattern, independent of its innateness, and that LMs dramatically expand our ability to study such patterns in silico. My proposal is that, by explicitly foregrounding algorithms as a distinguished class of real patterns – understood as interventionally robust causal models under causal abstraction – we can further consolidate the bridge between LMs and theoretical linguistics. LMs let us test, in unprecedented detail, which algorithms statistical learners converge on to solve linguistic tasks.

Financial support. Raphaël Millière is funded by an AI2050 Early Career Fellowship from Schmidt Sciences.

Competing interests. None.

References

Geiger, A., Harding, J., & Icard, T. (2025). *How Causal Abstraction Underpins Computational Explanation* (No. arXiv:2508.11214). arXiv. https://doi.org/10.48550/arXiv.2508.11214

Geiger, A., Ibeling, D., Zur, A., Chaudhary, M., Chauhan, S., Huang, J., Arora, A., Wu, Z., Goodman, N., Potts, C., & Icard, T. (2025). Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, *26*(83), 1–64.

Millière, R., & Rathkopf, C. (2026). Anthropocentric bias in language model evaluation. *Computational Linguistics*, *52*(1), 379–388. https://doi.org/10.1162/COLI.a.582

Piccinini, G., & Craver, C. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese*, *183*(3), 283–311. https://doi.org/10.1007/s11229-011-9898-4

Williams, D. J. (2025). Two senses of medium independence. *Mind & Language*, *40*(4), 437–447. https://doi.org/10.1111/mila.12542

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

Machine yearning: LLMs do not capture formal linguistic structure and obscure neuroscientific inquiry

Elliot Murphy[a]*, Paolo Morosi[b], Evelina Leivada[b,c] and Andrew Nevins[d]

[a]Vivian L. Smith Department of Neurosurgery, University of Texas Health Science Center at Houston, Houston, USA; [b]Universitat Autonoma de Barcelona, Spain; [c]Institució Catalana de Recerca i Estudis Avançats (ICREA), Spain and [d]Department of Linguistics, University College London, UK
elliot.murphy@uth.tmc.edu
paolo.morosi@uab.cat
evelina.leivada@uab.cat
a.nevins@ucl.ac.uk

*Corresponding author.

doi:10.1017/S0140525X26104634, e217

Abstract

Futrell and Mahowald argue that neural networks have learned non-trivial aspects of language. We argue that these systems have not in fact demonstrated "mastery" of syntax, marshalling recent evidence, and that they further obscure explanatory insights with respect to topics in the cognitive neuroscience of language.

Futrell and Mahowald (FM) argue that large language models (LLMs) have mastered the formal properties of human language. They point to work on the geometry of word embeddings and their ability to capture syntactic dependencies (Diego-Simón et al., 2024). However, dependency grammar is known to be not a viable candidate theory of natural language syntax, since it isolates word-word dependency graphs rather than hierarchical constituency structure (Marcolli, Chomsky, & Berwick, 2025) (dependency graphs are projections from this deeper algebraic structure). While FM (2025: 42) argue that recent work serves to "refute the claim that neural LMs can learn any language" (i.e., Kallini et al., 2024), this work is confounded by the fact that Kallini and colleagues did not contrast count-based vs. constituency-based non-adjacency rules. Instead, they contrasted count-based non-adjacency with simple adjacency, which is not the appropriate resolution to isolate constituency-based rules that are not operationalized over linear distance (Hunter 2024). Other work casts doubt on the claim that LLM probabilities can distinguish between possible/impossible languages: Leivada et al. (2025a) show that LLM probabilities do not constitute reliable proxies for model-internal representations of syntactic knowledge, and Yang et al. (2025) show that perplexity scores from GPT-2-small do not distinguish between attested and unattested word orders.

Natural language syntax-semantics is not exhausted in formal and conceptual scope by distributional similarity or predictive probability. Many words are able to jointly host analog/continuous and digital representations (Jackendoff & Erk, 2025), alongside categorially incompatible types of inherent polysemy (Murphy, 2024a) (consider what kind of thing *school* is meant to be in "The progressive school next to the river was high in the *NYT* rankings and had just hired a new chancellor after being repainted"). In contrast to vector space semantic models operating within a single homogeneous metric space, natural language semantics is highly typed and heterogeneous: the cognitive architecture underlying it has to coordinate multiple representational formats (spatial structure, event structure, theory of mind, quantification, Boolean logic, causation, telicity, and evidentiality). These conceptual domains are not "dimensions" in the embedding sense. They host different mathematical structures with different neural substrates (Baggio, 2018; Murphy, 2025).

If FM are right that LLMs have learned non-trivial aspects of language, what have they learned when they affirm that "Eight eighteen resigned seventeen this eight twenty-one any ten warning" is a grammatical sentence (Leivada et al., 2025b)? The field of artificial intelligence has often leaned on metaphors to circumvent the black box nature of its agents (Mitchell, 2024), with "learning" a clear example. The notion of emergence found in FM is another. Does any putative "mastery" of syntax entail that LLMs emulate the process of emergence as understood in biology (i.e., a behavior arising de novo as the (by-)product of a complex system)? Emergence is notoriously hard to define, but there is agreement that it entails novelty that is not predicted by lower-level generic components of the adaptive system (Krakauer, Krakauer, & Mitchell, 2025). Human-engineered systems do not straightforwardly satisfy standard criteria for strong emergence, because they are purposely designed to emulate target behavior. In this context, anthropomorphizing labels such as *learning*, *emerging*, and *reasoning* are more accurately described as wishful mnemonics (McDermott, 1981) – a textbook case of machine yearning – than reasonable scientific claims documenting model capabilities.

LLMs can surely capture certain statistical regularities from the output of language (text, data), but they do not infer grammar the way humans do and fail to capture fundamental principles of linguistic structure (Murphy et al., 2025). They do not reliably identify grammaticality (Dentella et al., 2024; Dentella, Günther, & Leivada, 2025a). Instead of performing higher-order reasoning, they have been shown to engage in "sophisticated pattern matching" (Shojaee et al., 2025). They have a strong linearity bias and do not capture deep syntactic structures (Diego-Simón et al., 2025). There is no evidence that LLMs capture cross-constructional generalizations about null complementizers (Momma, Richards, & Ferreira, 2025). When tested to learn restrictions on filler-gap dependencies in English and Norwegian, the generalizations learned by language models diverge from those of humans (Kobzeva et al., 2025). This does not support the claim that filler-gap dependencies and island constraints can be learned without domain-specific biases. LLMs have a preference for manipulating noun-related information over verb-related information (Dentella et al., 2025b), thus distinguishing them from highly grammaticalized types of thought that emphasize dimensions of tense, aspect, and mood. Lastly, text-to-image models also fail to represent the meaning of prompts with relatively basic syntactic structures triggering compositional meanings (Leivada, Murphy, & Marcus, 2023; Murphy, de Villiers, & Morales, 2025).

Within cognitive neuroscience, LLMs have arguably been unable to help with causal-mechanistic questions concerning *how* the algebraic properties of syntax-semantics are neurally enforced (Marcus, 2001; Murphy, 2015, 2020, 2024b, 2025; Murphy et al., 2022). Many LLM-brain alignments seem to be "driven by fragile methodologies and overlooked confounds" (e.g., word rate and positional signals) (Hadidi et al., 2025). An fMRI study showed that transformers were significantly inferior at predicting neural

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

signatures to models explicitly designed to encode the syntactic relations between words (Fodor, Murawski, & Suzuki, 2025). Problematic causal logic is also common in the LLM literature (e.g., affirming the consequent) (Guest & Martin, 2023), whereby researchers often confuse types of inference and misunderstand the evidentiary role that correlation plays (see also Meijer, 2021). A scientist or artificial model may be able to classify and predict very well – all without providing causal *explanations* of a system.

FM (2025: 3) warn against dismissing LLMs as approximation devices that can fit any data without revealing the true underlying system. But this analogy is apt. LLMs can mimic many linguistic phenomena because they are universal function approximators, and their capacity and data exposure are unbounded. LLM fluency should not be confused with competence (Slaats & Martin, 2025). We therefore stress that the type of functionalist approach to language discussed by FM sidelines the necessary contributions to theories of cognition from symbolic, algebraic structures.

FM invoke Kubrick's *Dr. Strangelove* in their article's title, referring to an emerging "love" for LLMs winning over irrational fear and anxiety. We feel that another Kubrick movie, *2001: A Space Odyssey*, teaches an altogether simpler and more appropriate lesson: to hesitate before trusting HAL's words.

Financial support. EL acknowledges funding from the Spanish Ministry of Science, Innovation & Universities (MCIN/AEI/https://doi.org/10.13039/501100011033) under the research project CNS2023-144415. The funders had no role in the decision to publish or the preparation of the manuscript.

Competing interests. None.

References

Baggio, G. (2018). *Meaning in the brain*. MIT Press.

Dentella, V., Günther, F., & Leivada, E. (2025a). Language in vivo vs. in silico: Size matters but larger language models still do not comprehend language on a par with humans due to impenetrable semantic reference. *PLoS One*, *20*(7), e0327794.

Dentella, V., Günther, F., Murphy, E., Marcus, G., & Leivada, E. (2024). Testing AI on language comprehension tasks reveals insensitivity to underlying meaning. *Scientific Reports*, *14*, 28083.

Dentella, V., Huang, W., Mansi, S. A., Grieve, J., & Leivada, E. (2025b). ChatGPT–generated texts show authorship traits that identify them as non-human. arXiv:2508.16385.

Diego-Simón, P., D'Ascoli, S., Chemla, E., Lakretz, Y., & King, J. R. (2024). A polar coordinate system represents syntax in large language models. In *38th Conference on Neural Information Processing Systems (NeurIPS 2024)*, Vancouver, Canada.

Diego-Simón, P. J., Chemla, E., King, J. R., & Lakretz, Y. (2025). Probing syntax in large language models: Successes and remaining challenges. arXiv:2508.03211.

Fodor, J., Murawski, C., & Suzuki, S. (2025). When word order matters: Human brains represent sentence meaning differently from large language models. *eLife*, *14*, RP108442.

Guest, O., & Martin, A. E. (2023). On logical inference over brains, behaviour, and artificial neural networks. *Computational Brain & Behavior*, *6*, 213–227.

Hadidi, N., Feghhi, E., Song, B. H., Blank, I. A., & Kao, J. C. (2025). Illusions of alignment between large language models and brains emerge from fragile methods and overlooked confounds. bioRxiv https://doi.org/10.1101/2025.03.09.642245.

Hunter, T. (2024). Kallini et al. (2024) do not compare impossible languages with constituency-based ones. arXiv:2410.12271.

Jackendoff, R., & Erk, K. E. (2025). Toward a deeper lexical semantics. *Topics in Cognitive Science*, *17*(4), 962–972. https://doi.org/10.1111/tops.70013.

Kallini, J., Papadimitriou, I., Futrell, R., Mahowald, K., & Potts, C. (2024). Mission: Impossible language models. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 14691–14714. https://doi.org/10.18653/v1/2024.acl-long.787

Kobzeva, A., Arehalli, S., Linzen, T., & Kush, D. (2025). Learning filler-gap dependencies with neural language models: Testing island sensitivity in Norwegian and English. *Journal of Memory and Language*, *144*, 104663.

Krakauer, D. C., Krakauer, J. W., & Mitchell, M. (2025). Large language models and emergence: A complex systems perspective. arXiv:2506.11135.

Leivada, E., Marcus, G., Günther, F., & Murphy, E. (2025b). A sentence is worth a thousand pictures: Can Large Language Models understand hum4n l4ngu4ge and the w0rld behind w0rds? *Philosophical Transactions of the Royal Society A* (Forthcoming).

Leivada, E., Montero, R., Morosi, P., Moskvina, N., Serrano, T., Aguilar, M., & Guenther, F. (2025a). Large language model probabilities cannot distinguish between possible and impossible languages. arXiv:2509.15114.

Leivada, E., Murphy, E., & Marcus, G. (2023). DALL·E 2 fails to reliably capture common syntactic processes. *Social Sciences & Humanities Open*, *8*(1), 100648.

Marcolli, M., Chomsky, N., & Berwick, R. C. (2025). *Mathematical structure of syntactic merge: An algebraic model for generative linguistics*. MIT Press.

Marcus, G. F. (2001). *The algebraic mind: Integrating connectionism and cognitive science*. MIT Press.

McDermott, D. (1981). Artificial Intelligence meets natural stupidity. *ACM SIGART Bulletin*, *57*, 4–9.

Meijer, G. (2021). Neurons in the mouse brain correlate with cryptocurrency price: A cautionary tale. *Peer Community Journal*, *1*, e29.

Mitchell, M. (2024). The metaphors of artificial intelligence. *Science*, *386*(6723), eadt6140.

Momma, S., Richards, N., & Ferreira, V. S. (2025). Speakers encode silent structures: Evidence from complementizer priming in English. *Journal of Memory and Language*, *145*, 104671.

Murphy, E. (2015). The brain dynamics of linguistic computation. *Frontiers in Psychology*, *6*, 1515.

Murphy, E. (2020). *The oscillatory nature of language*. Cambridge University Press.

Murphy, E. (2024a). Predicate order and coherence in copredication. *Inquiry*, *67*(6), 1744–1780.

Murphy, E. (2024b). ROSE: A neurocomputational architecture for syntax. *Journal of Neurolinguistics*, *70*, 101180.

Murphy, E. (2025). ROSE: A universal neural grammar. *Cognitive Neuroscience*, *16*(1–4), 49–80.

Murphy, E., de Villiers, J., & Morales, S. L. (2025). A comparative investigation of compositional syntax and semantics in DALL·E and young children. *Social Sciences & Humanities Open*, *11*, 101332.

Murphy, E., Leivada, E., Dentella, V., Günther, F., & Marcus, G. (2025). Fundamental principles of linguistic structure are not represented by ChatGPT. *Biolinguistics*, *19*, 1–55.

Murphy, E., Woolnough, O., Rollo, P. S., Roccaforte, Z., Segaert, K., Hagoort, P., & Tandon, N. (2022). Minimal phrase composition revealed by intracranial recordings. *Journal of Neuroscience*, *42*(15), 3216–3227.

Shojaee, P., Mirzadeh, I., Alizadeh, K., Horton, M., Bengio, S., & Farajtabar, M. (2025). The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. Apple Machine Learning Research report. June.

Slaats, S., & Martin, A. E. (2025). What's surprising about surprisal. *Computational Brain & Behavior*, *8*, 233–248.

Yang, X., Aoyama, T., Yao, Y., & Wilcox, E. (2025). Anything goes? A cross-linguistic study of (im)possible language learning in LMs. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 26058–26077.

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

Across the levels of analysis: Explaining predictive processing in humans requires more than machine-estimated probabilities

Sathvik Nair[a]* and Colin Phillips[a,b]

[a]University of Maryland, College Park: University of Maryland, USA and [b]Oxford University: University of Oxford, UK
sathvik@umd.edu
colin.phillips@ling-phil.ox.ac.uk

*Corresponding author.

doi:10.1017/S0140525X26104701, e218

Abstract

Under the lens of Marr's (levels of analysis, we critique and extend the authors' two points about language models (LMs) and language processing: first, predicting upcoming linguistic information based on context is key to language processing, and second, that many advances in psycholinguistics would be impossible without LLMs. We also outline directions combining LLMs' strengths with psycholinguistic models.

In the 1960s, psycholinguists were excited about the possibility of a **computational-level** account of language processing where measures of "processing difficulty" were linked to aggregate measures of linguistic complexity. In that case, the difficulty measure (ratings or reading times for whole sentences) and the complexity measure (syntactic transformations) reflected the available (psycho)linguistic technologies of the time (Miller & McKean, 1964). The so-called Derivational Theory of Complexity (DTC) is widely cited as a cautionary tale about What Not to Do in cognitive science, but there are some enduring results establishing expectations for relative processing difficulty among different syntactic constructions.

Fifty to sixty years later, we again find enthusiasm for a computational-level account that links overall measures of processing difficulty to aggregate measures of linguistic complexity. Although today's technical advances reflect richer processing data (eye fixation durations, electrical potentials in the brain, or utterance latencies) and linguistic complexity (contextual probabilities estimated by LLMs), all over words (and subword units) instead of sentences, the core idea retains the half-century-old notion that a general-purpose measure to explain processing difficulty is a useful construct, now expressed as surprisal theory (Levy, 2008). That said, it is important to learn from the mistakes of the DTC.

Just as in the 1960s, there are more cases where the general theory falls short, but there are some core insights that the field is reluctant to abandon. In the 2020s, these insights pertain to predictability effects in language processing (Staub, 2025); unlikely, information in a given context is more difficult to process. Even simple LMs quantitatively describe these effects well, which has led researchers to claim language processing is associated with probabilistic inference (Smith & Levy, 2013). LLMs have since characterized predictability effects across different languages (Wilcox et al., 2023; Xu et al., 2023) and larger datasets (Shain et al., 2024), but are largely based on the same arguments. A major issue in this debate is that for LMs, predictability effects are *exclusively* linked to probabilistic inference, while the same cannot necessarily be said for humans, thanks to insights from the last few decades of psycholinguistic research (Brothers & Kuperberg, 2021; Szewczyk & Federmeier, 2022).

From the fall of the DTC to the rise of surprisal theory, psycholinguists discovered finer-grained phenomena and experimental measures, largely shifting their attention to **algorithmic-level** questions. This fueled decades of research, and so few considered psycholinguistics to be equipped to make computational-level contributions. Futrell & Mahowald do say language processing is highly incremental, but psycholinguists' goal has been seen as characterizing the *underlying mental computations* from each step of incremental processing.

Contrary to the authors' claims that psycholinguistics has been hampered by weak statistical models of language, the field has made progress through empirical investigations of phenomena testing the processing system in tightly controlled experimental settings. Some examples include studies on linguistic illusions (Paape, 2024; Phillips, Wagers, & Lau, 2011), consequences of violated predictions (Frisson, Harvey, & Staub, 2017; Kuperberg, Brothers, & Wlotko, 2020), time delays (Chow et al., 2018; Tabor & Hutchins, 2004), and production-comprehension contrasts (Kandel, Wyatt, & Phillips, 2024; Lee & Phillips, 2025). Simpler, interpretable models, not LLMs, have proven useful in this context (Aurnhammer et al., 2021; Fitz & Chang, 2019; Nakamura et al., 2024).

That said, an important theoretical contribution of LLMs is to highlight the limits of predictability as a satisfying explanation for finer-grained aspects of language processing that unfold in time. LLM probabilities have limits as a functional explanation of human neural responses because they conflate semantic association with contextual expectations (Krieger et al., 2025). On the behavioral side, they insufficiently capture qualitative patterns in argument role reversals (Lee, Nair, & Feldman, 2024) and other types of linguistic illusions (Zhang et al., 2023), and quantitative patterns in syntactic processing (Huang et al., 2024).

These recent findings mark a shift from computational-level probabilities back towards algorithmic-level processes. To this end, Futrell & Mahowald suggest imposing cognitively realistic resource constraints on LLMs' predictability estimates, and Giulianelli, Wallbridge, and Fernández (2023, 2024) derive LLM-based difficulty measures for different levels of linguistic representation. We suggest another direction for progress: building on interactivity across levels of representation and incorporating predictability-based factors in process models. The field is moving in this direction already; LLMs complement insights about the time course of syntactic processing (Slaats, Meyer, & Martin, 2024; Timkey et al., 2025), can be combined with syntactic and semantic measures (De Santo, 2025; Jacobs et al., 2025; Stanojević et al., 2023), and support algorithmic-level accounts of inference (Vigly et al., 2025).

Process models can also be motivated by **implementational-level** ideas such as predictive coding (Clark, 2013; Friston, 2009; Rao & Ballard, 1999), grounded in theories of neurobiological function. Even if Futrell & Mahowald cite this work, LLMs' successes are at best rhetorically connected with predictive coding due to their limited biological plausibility (Goldstein et al., 2022; Ryskin & Nieuwland, 2023; Schrimpf et al., 2021).

Building on successfully implemented predictive coding models in other cognitive domains, their recent extension to linguistic data is promising. Not only do they elegantly explain disparate effects in

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

controlled experimental settings (Nour Eddine et al., 2024), they also explain naturalistic reading data, with predictive power comparable to LMs with similar size (Ohams et al., 2026). Crucially, both these advances were derived from connecting simpler cognitive models (Elman, 1991; Rumelhart & McClelland, 1981) to biologically plausible methods of computation and provide interpretable measures that are directly tied to interactive cognitive processes, unlike LLMs.

To summarize, LMs have "revived" an older psycholinguistic approach, now providing far more accurate estimates of aggregate processing difficulty and linguistic complexity. Probabilistic models of linguistic data are useful, but progress in psycholinguistics will also require mechanistically characterizing the mental computations at the algorithmic level, in order to translate estimates of *which* words are (un)predictable from above into neural processes of *how* they are processed from below. Thus, psycholinguists can find different ways of loving what LMs can do for them, either as models of the core mechanism in language processing or as components of more articulated algorithms, but their worries are justified.

Acknowledgements. We are grateful to Samer Nour Eddine, Philip Resnik, Ziying Zhang, Matt Husband, and Alba Jorquera Jiménez de Aberásturi for valuable discussions.

Financial support. Sathvik Nair is supported by a NSF GRFP under Grant No. DGE 2236417.

Competing interests. None to declare.

References

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbauam, J.B, & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, *118*(45), e2105646118.

Aurnhammer, C., Delogu, F., Schulz, M., Brouwer, H., & Crocker, M. W. (2021). Retrieval (N400) and integration (P600) in expectation-based comprehension. *Plos one*, *16*(9), e0257430.

Brothers, T., & Kuperberg, G. R. (2021). Word predictability effects are linear, not logarithmic: Implications for probabilistic models of sentence comprehension. *Journal of Memory and Language*, *116*, 104174.

Chow, W. Y., Lau, E., Wang, S., & Phillips, C. (2018). Wait a second! Delayed impact of argument roles on on-line verb prediction. *Language, Cognition and Neuroscience*, *33*(7), 803–828.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204.

De Santo, A. (2025). Capturing online SRC/ORC effort with memory measures from a minimalist parser. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics* (pp. 24–35).

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine learning*, *7*(2), 195–225.

Fitz, H., & Chang, F. (2019). Language ERPs reflect learning through prediction error propagation. *Cognitive Psychology*, *111*, 15–52.

Frisson, S., Harvey, D. R., & Staub, A. (2017). No prediction error cost in reading: Evidence from eye movements. *Journal of Memory and Language*, *95*, 200–214.

Friston, K. (2009). The free-energy principle: A rough guide to the brain?. *Trends in Cognitive Sciences*, *13*(7), 293–301.

Giulianelli, M., Wallbridge, S., Cotterell, R., & Fernández, R. (2024). Incremental alternative sampling as a lens into the temporal and representational resolution of linguistic prediction. *PsyArXiv*.

Giulianelli, M., Wallbridge, S., & Fernández, R. (2023). Information value: Measuring utterance predictability as distance from plausible alternatives. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 5633–5653).

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., et al. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, *25*(3), 369–380.

Huang, K. J., Arehalli, S., Kugemoto, M., Muxica, C., Prasad, G., Dillon, B., & Linzen, T. (2024). Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language*, *137*, 104510.

Jacobs, C. L., Hubbard, R. J., Grobol, L., & Federmeier, K. D. (2025). Uncovering patterns of semantic predictability in sentence processing. *Journal of Memory and Language*, *144*, 104653.

Kandel, M., Wyatt, C. R., & Phillips, C. (2024). Number attraction in pronoun production. *Open Mind*, *8*, 1247–1290.

Krieger, B., Brouwer, H., Aurnhammer, C., & Crocker, M. W. (2025). On the limits of LLM surprisal as a functional explanation of the N400 and P600. *Brain Research*, *1865*, 149841.

Kuperberg, G. R., Brothers, T., & Wlotko, E. W. (2020). A tale of two positivities and the N400: Distinct neural signatures are evoked by confirmed and violated predictions at different levels of representation. *Journal of Cognitive Neuroscience*, *32*(1), 12–35.

Lee, E. K., Nair, S., & Feldman, N. (2024). A psycholinguistic evaluation of language models' sensitivity to argument roles. In *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 3262–3274).

Lee, E. K. R., & Phillips, C. (2025). Argument role sensitivity in real-time sentence processing: Evidence from a hybrid comprehension and production task. *Cognition*, *264*, 106255.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W.H. Freeman.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological review*, *88*(5), 375.

Miller, G. A., & McKean, K. O. (1964). A chronometric study of some relations between sentences. *Quarterly Journal of Experimental Psychology*, *16*(4), 297–308.

Nakamura, M., Momma, S., Sakai, H., & Phillips, C. (2024). Task and timing effects in argument role sensitivity: Evidence from production, EEG, and computational modeling. *Cognitive Science*, *48*(12), e70023.

Nour Eddine, S. N., Brothers, T., Wang, L., Spratling, M., & Kuperberg, G. R. (2024). A predictive coding model of the N400. *Cognition*, *246*, 105755.

Ohams, C., Nair, S., Bhattasali, S., & Resnik, P. (2026). A predictive coding model for online sentence processing. *Journal of Memory and Language*, *146*, 104705.

Paape, D. (2024). How do linguistic illusions arise? Rational inference and good-enough processing as competing latent processes within individuals. *Language, Cognition and Neuroscience*, *39*(10), 1334–1365.

Phillips, C., Wagers, M. W., & Lau, E. F. (2011). Grammatical illusions and selective fallibility in real-time language comprehension. *Experiments at the Interfaces*, *37*, 147–180.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87.

Ryskin, R., & Nieuwland, M. S. (2023). Prediction during language comprehension: What is next?. *Trends in Cognitive Sciences*, *27*(11), 1032–1052.

Shain, C., Meister, C., Pimentel, T., Cotterell, R., & Levy, R. (2024). Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, *121*(10), e2307876121.

Slaats, S., Meyer, A. S., & Martin, A. E. (2024). Lexical surprisal shapes the time course of syntactic structure building. *Neurobiology of Language*, *5*(4), 942–980.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319.

Stanojević, M., Brennan, J. R., Dunagan, D., Steedman, M., & Hale, J. T. (2023). Modeling structure-building in the brain with CCG parsing and large language models. *Cognitive science*, *47*(7), e13312.

Staub, A. (2025). Predictability in language comprehension: Prospects and problems for surprisal. *Annual Review of Linguistics*, *11*(1), 17–34.

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

Szewczyk, J. M., & Federmeier, K. D. (2022). Context-based facilitation of semantic access follows both logarithmic and linear functions of stimulus probability. *Journal of memory and language*, *123*, 104311.
Tabor, W., & Hutchins, S. (2004). Evidence for self-organized sentence processing: Digging-in effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(2), 431.
Timkey, W., Huang, K. J., Oh, B. D., Prasad, G., Arehalli, S., Linzen, T., & Dillon, B. (2025). Eye movements reveal a dissociation between prediction and structural processing in language comprehension. *Preprint.*
Vigly, J., Qian, P., Sonderegger, M., & O'Donnell, T. (2025). Comprehension effort as the cost of inference. *PsyArxiv*.
Wilcox, E. G., Pimentel, T., Meister, C., Cotterell, R., & Levy, R. P. (2023). Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, *11*, 1451–1470.
Xu, W., Chon, J., Liu, T., & Futrell, R. (2023). The linearity of the effect of surprisal on reading times across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 15711–15721).
Zhang, Y., Gibson, E., & Davis, F. (2023). Can language models be tricked by language illusions? Easier with syntax, harder with semantics. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)* (pp. 1–14).

The language of real patterns

Ryan Mark Nefdt[a]* and James Ladyman[b]

[a]Philosophy, University of Cape Town, Cape Town, South Africa and [b]Philosophy, University of Bristol, UK
ryan.nefdt@uct.ac.za
james.ladyman@bristol.ac.uk

*Corresponding author.

doi:10.1017/S0140525X26104580, e219

Abstract

In this response, we outline the central idea of a real pattern as applied in the philosophy of science as well as its relationship to the modal structure of models. We suggest that there may be additional concerns with linguistically real patterns in terms of convergence between natural language processing and formal linguistic theory.

Dennett's idea of real patterns connects emergent ontology with projectible regularities in the phenomena. In Dennett's example of the *Game of Life*, there are emergent "entities" (or patterns), such as eaters and gliders, whose behaviour is predictable. In human behaviour, there are real patterns too in the form of the entities of folk psychology – for example, beliefs and desires – and according to Dennett, the question of whether such things exist is exhausted by the question of whether there are explanatorily and predictively successful causal or lawlike regularities expressed in terms of them (in which they play a non-redundant role). Dennett's idea has been applied much more widely recently, for example, by David Wallace to the emergent ontology of fluid mechanics (Wallace, 2026), and Margarida Hermida to biological classification (Hermida, 2026). It has been applied to linguistic phenomena by Ryan Nefdt (2023). The key aspect of real patterns for present purposes is that they depend on there being a modal structure to the relevant phenomena that is represented by a scientific theory or model. Philosophers of science have studied scientific modelling of many kinds in great depth, and there are two important lessons that have been learned pertinent to the debate about large language models and natural language. The first is that models can represent modal structure approximately and incompletely, without there being an isomorphism between the structure of the domain and the model. The second, which is a corollary of the first, is that there can be multiple models of the same domain that represent different aspects of its modal structure.

Futrell and Mahowald ingeniously apply the idea of real patterns to deep learning language models (LMs), prominent in current natural language processing (NLP). They argue that these models can reveal linguistic structure via simple compressed representations of past regularities found in natural language data. Their account aims to circumnavigate the charybdis of formal linguistics, which has persistently rejected the importance of statistical methods, and the scylla of modern NLP, which shows an inclination to interpret the successes of LMs as refutations of said formal linguistic theory. We commend their effort, but one important question remains unaddressed namely, what is the domain of these real patterns? In particular, are they patterns in linguistic outputs or are they patterns in cognition?

Formal linguists, following Chomsky, emphasise the cognitive-explanatory nature of the science above all else. If there are real patterns to be found, they likely lurk surreptitiously in the mind of cognisers. Compressibility will then require formal, discrete mathematical tools (like proof-theoretic generative grammars). In contrast, computational linguistics has always been performance-based, and in fact, Futrell and Mahowald do see LMs as vindications of usage-based theories of language. Here, big data, statistics and more recently, deep neural networks reveal structure, of which the categories of linguistic theory might only be useful and lossy representations. But Futrell and Mahowald seem to see the Dennettian picture as that of linguistic structure (e.g. "grammatical subject") being at the level of belief or folk psychology and the messy underlying circuitry living at the level of neural nets, big data, and statistics. However, if we return to the Game of Life, it seems to be the other way around, the underlying reality is characterised by simple discrete rules (like a cell stays alive if it has 2 or 3 live neighbours on a 2D grid), and the emergent complexity of gliders, eaters, and pulsars is to be found at higher levels. It marks a genuine phase transition. If this analogy applies, then their argument is reversed. Formal linguistic structures are not necessarily emergent real patterns but deeper characterisations of a deterministic universe. This actually corresponds with the idea that LMs are prediction machines. What makes intentional vocabulary (beliefs, desires, intentions) useful and ineliminable (for Dennett) is that it provides us with mechanisms for the successful prediction of behaviour. The formal theory furnishes the fundamental physics. But here too there is a caveat since LMs often do not provide their predications in a human transparent manner (although perhaps neither do brains, see Hasson, Nastase, & Goldstein, 2020), and formal linguistics, unlike physics, does not always deliver on the prediction or implementation side.

Of course, it is possible that nature endowed human beings with a redundant linguistic architecture such that both formal grammar rules and the statistical approximations of LMs can represent linguistic structure to different degrees. This would obviate the need to find the actual or underlying modal base of linguistic structure (i.e. is natural language discrete or continuous in essence?). Competence and performance would be bound up in a complex system. Perhaps, as F&M suggest, LMs can serve as proofs of concept or delimiters of the modal space of possible linguistic

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

structure, within which human representational capabilities operate. It really depends on what kinds of models LMs are. It is also possible that, with relation to language, we simply have "multiple models idealisation" (Weisberg, 2013), similar to modelling practices in sciences like climate science. Multiple incompatible representations and resources track different patterns in accordance with distinct scientific purposes and goals. Without a clear path to connecting the two paradigms in linguistic theorising, it seems unclear how real patterns can unify or unite formal linguistics with NLP and engineering.

This all comes back to our initial question: where do the real patterns reside? Initially, both formal symbolic grammar and the subsymbolic nature of connectionism were searching for them within cognition, such that classic proposals like Smolensky (1986) aim at reconciliation within one framework. Or perhaps something akin to autoregression is at the heart of human language competence as well (Caucheteux & King, 2022)? Either way, we have to contend with the fact that modern LMs are not designed to explain or uncover cognitive structure explicitly; they are engineering devices created to optimise a next token prediction task. It remains to be seen if the patterns they so expertly and superhumanly survey are the *real* ones in the heads of language users.

Financial support. No funding to report.

Competing interests. There are no competing interests.

References

Caucheteux, C., & King, J. R. (2022). Brains and algorithms partially converge in natural language processing. *Communications Biology*, *5*, 134. https://doi.org/10.1038/s42003-022-03036-1

Hasson, U., Nastase, S. A., & Goldstein, A. (2020). Direct fit to nature: An evolutionary perspective on biological and artificial neural networks. *Neuron*, *105*(3), 416–434. https://doi.org/10.1016/j.neuron.2019.12.002

Hermida, M. (2026). Disinterested taxonomic pluralism: Biological classifications track real patterns. In T. Millhouse, S. Petersen, & D. Ross (Eds.), *Real patterns in science and nature* (Accepted/In press). MIT Press.

Nefdt, R. (2023). *Language, science, and structure*. Oxford University Press.

Smolensky, P. (1986). Neural and conceptual interpretation of PDP models. In *Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 2: Psychological and biological models* (pp. 390–431). MIT Press.

Wallace, D. (2026). Real pattern in physics and beyond. In T. Millhouse, S. Petersen, & D. Ross (Eds.), *Real patterns in science and nature* (Accepted/In press). MIT Press.

Weisberg, M. (2013). *Simulation and similarity*. Oxford University Press.

Are language models models?

Philip Resnik*

Department of Linguistics and Institute for Advanced Computer Studies, University of Maryland, College Park, USA
resnik@umd.edu

*Corresponding author.

doi:10.1017/S0140525X26104762, e220

Abstract

Futrell and Mahowald claim language models (LMs) "serve as model systems," but an assessment at each of Marr's three levels suggests the claim is clearly not true at the implementation level, poorly motivated at the algorithmic-representational level, and problematic at the computational theory level. LMs are good candidates as tools; calling them cognitive models overstates the case and unnecessarily feeds large language model hype.

It's impossible to resist pointing out that in Stanley Kubrick's 1964 film *Dr. Strangelove*, which inspired Futrell and Mahowald's article title, the central theme involved people not thinking things through carefully enough, leading to a global apocalypse. I'm not worried about quite *that* level of catastrophe, mind you, but I do think the article risks further feeding into a pervasive and ultimately harmful misperception about the relationship between language models (LMs) and human language processing – namely, the too-readily overextended idea of LMs as "model systems" alluded to in the article's abstract.

What do we mean when we say "model system"? A model, at any level of explanation, is an embodiment of theory that, in addition to capturing generalizations and licensing verifiable inferences and predictions – requirements of any theory – also contains structured correspondences between entities and relationships in the model and corresponding entities and relationships in the real-world system being modeled (Frigg & Nguyen, 2017). Notwithstanding its imperfections, Marr (1982) remains a widely accepted scaffolding (Peebles & Cooper, 2015), so let's consider the status of and prospects for LMs as models of language processing at each of Marr's three levels.

First, the **implementation level,** which we need not spend much time on. The authors argue for a hypothesis-generation role, but they are also clearly aware that functional parallels are not implementation-level models. Indeed, the connectionism underlying today's LMs was founded on the *abandonment* of implementation-level biological correspondences in favor of function optimization. In Fridman (2021), Jay McClelland recalls Geoff Hinton successfully arguing, "The problem with the way you guys have been approaching this is that you've been looking for inspiration from biology . . . That's the wrong way to go about it." Consequently, the dominant LM research today demonstrates little substantive engagement with the numerous relevant advances made in implementation-level understanding of the brain since connectionism's loose 20th-century inspirations (e.g., Barrett & Satpute, 2013; Hawkins & Ahmad, 2016; Hickok & Poeppel, 2007; Wright, Hedrick, & Komiyama, 2025; for higher-level reviews see Pulvermüller et al., 2021; Pulvermüller, 2023 and Durstewitz, Averbeck, & Koppe, 2025).

Algorithmic/representational level. The authors' argument here hinges on convergence, that systems solving the same hard problem are driven toward similar mechanisms (Cao & Yamins, 2024). But unlike visual object recognition, a key part of what makes "language" hard is that it decomposes into very distinct subproblems, not merely competing constraints within a single problem. In addition, mechanistic exceptions to convergence are easily found in biology (e.g., sharks and dolphins facing ecologically similar prey-capture problems evolved markedly different mechanisms, relying on electroreception and echolocation,

respectively) and in computation (e.g., iterative and recursive computations of the Fibonacci sequence). In medical research, model-organism choices are supported by existing, solidly established mechanistic evidence; e.g., mice and ferrets are preferred in work on tumorigenesis and influenza, respectively – not vice versa – because mice share oncogene pathways with humans and ferrets share lung physiology (Belser, Katz, & Tumpey, 2011; Bedell, Jenkins, & Copeland, 1997).

Consider the following scenario. Two aliens from another galaxy, whom I'll dub Abbot and Costello after Villeneuve (2016), show up on Earth with their own billions of years of evolution behind them. In a year, they absorb nearly everything ever written, and become phenomenal conversationalists; moreover, they generously offer to let human scientists experiment and probe their brain-boxes to our hearts' content. Here's the question: why in the world would one *expect* to find mechanistic correspondences between what's happening in Abbot and Costello's boxes and what is happening in the human cognitive apparatus? Nonetheless, for LMs, the "argument from amazingness" is pervasive – if they are doing something so very human so very well, surely they must be operating under the hood in human-like ways?

As I've highlighted, the argument is unreliable at best, and, moreover, it just is not necessary in order for LMs to contribute value to computational cognitive modeling. Giulianelli, Opedal, and Cotterell (2024), for example, use large language models to generate a sample of continuations in order to operationalize a generalization of surprisal that captures distinctions between anticipatory and responsive online processes in comprehension. This use depends only on the generation of plausible text, LMs' undisputed strength (for better or worse; see Resnik, 2025), without requiring claims about their cognitive plausibility. Similarly, LMs can provide useful proxies (with caveats; see Oh & Schuler, 2023) for human comprehenders' next-word probabilities when operationalizing models where such probabilities play a role, e.g., Hahn et al.'s (2022) integration of expectation- and memory-based theories of processing difficulty. LMs are a promising source of world knowledge for hierarchical predictive coding models, which are neurobiologically plausible and have recently been applied to language comprehension (Nour Eddine et al., 2024; Ohams et al., 2026).

Computational theory level. Ironically, many who "love the language models" seem to exhibit the same physics envy that you find in Chomskian theory, idealizing a uniform, elegant computation-level paradigm for a wide range of phenomena. But evolution is a scruffy tinkerer, not a neat software engineer (Jacob, 1977). It doesn't re-design from the bottom up when it produces new functionality. For LMs, just as for Chomsky, the drive to explain "du visible compliqué par de l'invisible simple" (Perrin, 1913) creates a problematic gap: even if these architectures can approximate diverse linguistic phenomena, and even if correlations are found between one black box and the other, ultimately there remains the burden of mapping any such computational theory to a biologically plausible algorithmic account of real-time processes distributed over diverse, specialized components. And there are plenty of other, less problematic computational-theory games in town (e.g., Eliasmith, 2013; Suárez et al., 2024; Sprevak & Smith, 2023).

At all levels, we do need a pluralistic approach, and LMs can be tremendously productive as *tools* in support of developing explanatory models. But LMs themselves are poor candidates as "model systems" at any of Marr's three levels. The distinction matters, for how we prioritize research questions, interpret results, and communicate findings to neighboring fields. For a whole variety of reasons, large language models are the subject of enormous hype. Let's not undermine some of the more nuanced arguments in this paper by overstating the case.

Acknowledgments. Helpful comments from Katherine Howitt, Colin Phillips, and Samer Nour Eddine were much appreciated.

Financial support. This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Competing interests. None.

References

Barrett, L. F., & Satpute, A. B. (2013). Large-scale brain networks in affective and social neuroscience: Towards an integrative functional architecture of the brain. *Current Opinion in Neurobiology*, *23*(3), 361–372.

Bedell, M. A., Jenkins, N. A., & Copeland, N. G. (1997). Mouse models of human disease. Part I: techniques and resources for genetic analysis in mice. *Genes & Development*, *11*(1), 1–10. https://doi.org/10.1101/gad.11.1.1

Belser, J. A., Katz, J. M., & Tumpey, T. M. (2011). The ferret as a model organism to study influenza A virus infection. *Disease Models & Mechanisms*, *4*(5), 575–579. https://doi.org/10.1242/dmm.007823

Cao, R., & Yamins, D. (2024). Explanatory models in neuroscience, Part 2: Functional intelligibility and the contravariance principle. *Cognitive Systems Research*, *85*, 101200.

Durstewitz, D., Averbeck, B., & Koppe, G. (2025). What neuroscience can tell AI about learning in continuously changing environments. *Nature Machine Intelligence*, *7*, 1897–1912. https://doi.org/10.1038/s42256-025-01146-z

Eliasmith, C. (2013). *How to build a brain: A neural architecture for biological cognition*. Oxford University Press.

Fridman, L. (2021, September 28). *Geoffrey Hinton is a genius | Jay McClelland and Lex Fridman* [Video]. YouTube. https://www.youtube.com/watch?v=lLBbsif2Xt4

Frigg, R., & Nguyen, J. (2017). Models and representation. In L. Magnani, & T. Bertolotti (Eds.), *Springer handbook of model-based science* (pp. 49–102). Springer International Publishing.

Giulianelli, M., Opedal, A., & Cotterell, R. (2024). Generalized measures of anticipation and responsivity in online language processing. In *Findings of the association for computational linguistics: EMNLP 2024* (pp. 11648–11669). Association for Computational Linguistics. https://aclanthology.org/2024.findings-emnlp.682/

Hahn, M., Futrell, R., Levy, R., & Gibson, E. (2022). A resource-rational model of human processing of recursive linguistic structure. *Proceedings of the National Academy of Sciences*, *119*(43), e2122602119.

Hawkins, J., & Ahmad, S. (2016). Why neurons have thousands of synapses, a theory of sequence memory in neocortex. *Frontiers in Neural Circuits*, *10*, 23.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*(5), 393–402.

Jacob, F. (1977). Evolution and tinkering. *Science*, *196*(4295), 1161–1166.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W. H. Freeman.

Nour Eddine, S., Brothers, T., Wang, L., Spratling, M., & Kuperberg, G. R. (2024). A predictive coding model of the N400. *Cognition*, *246*, 105755.

Oh, B., & Schuler, W. (2023). Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, *11*, 336–350.

Ohams, C., Nair, S., Bhattasali, S., & Resnik, P. (2026). A predictive coding model for online sentence processing. *Journal of Memory and Language*, *146*, 104705.

Peebles, D., & Cooper, R. P. (2015). Thirty years after Marr's vision: Levels of analysis in cognitive science. *Topics in Cognitive Science*, *7*(2), 187–190.

Perrin, J. (1913). *Les atomes*. Félix Alcan.

Pulvermüller, F., Tomasello, R., Henningsen-Schomers, M. R., & Wennekers, T. (2021). Biological constraints on neural network models of cognitive function. *Nature Reviews Neuroscience*, *22*, 488–502.

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

Pulvermüller, F. (2023). Neurobiological mechanisms for language, symbols and concepts: Clues from brain-constrained deep neural networks. *Progress in Neurobiology*, *230*, 102511.
Resnik, P. (2025). Large language models are biased because they are large language models. *Computational Linguistics*, *51*(3), 885–906.
Sprevak, M., & Smith, R. (2023). An introduction to predictive processing models of perception and decision-making. *Topics in Cognitive Science*. https://doi.org/10.1111/tops.12704
Suárez, L. E., Mihalik, A., Milisav, F., Marshall, K., Li, M., Vértes, P. E., Lajoie, G., & Misic, B. (2024). Connectome-based reservoir computing with the conn2res toolbox. *Nature Communications*, *15*(1), 656.
Villeneuve, D. (Director). (2016). *Arrival*. [Film]. Paramount Pictures.
Wright, W. J., Hedrick, N. G., & Komiyama, T. (2025). Distinct synaptic plasticity rules operate across dendritic compartments in vivo during learning. *Science*, *388*(6744), 322–328.

Large language models illuminate the mechanistic underpinnings of the creative aspect of language use (CALU), long regarded as a mystery

Chandra Sripada[a]* , Andrew McInnerney[b] and Richard L. Lewis[c]

[a]Psychiatry, Philosophy, Weinberg Institute for Cognitive Science, University of Michigan, Ann Arbor, USA; [b]Weinberg Institute for Cognitive Science, University of Michigan, Ann Arbor, USA and [c]Psychology, Linguistics, Weinberg Institute for Cognitive Science, University of Michigan, Ann Arbor, USA
sripada@umich.edu
amcin@umich.edu
rickl@umich.edu

*Corresponding author.

doi:10.1017/S0140525X26104725, e221

Abstract

Large language models (LLMs) challenge Chomsky's long-standing mysterian view of the creative aspect of language use (CALU). By exhibiting fluent, situation-appropriate linguistic behavior and offering concrete mechanistic hypotheses, they provide the first viable scientific models of CALU. We endorse Futrell and Mahowald's call to integrate LLMs into linguistic inquiry and suggest a bolder aim: elucidating the mechanisms underlying linguistic creativity.

Futrell and Mahowald argue convincingly that LLMs can expand linguistics, opening the field to neglected questions and creating bridges across disciplines. Their focus is on how LLMs can (potentially) explain the acquisition and implementation of syntactic regularities – a traditional focus of linguistics, especially in the Chomskian tradition. But why stop at syntax? We suggest a bolder aim in the same general spirit. LLMs can illuminate perhaps the deepest enigma about language: what Chomsky called the creative aspect of language use (CALU) (Chomsky, 2009).

CALU refers to speakers' ability to freely select, from the vast array of grammatically possible sentences, utterances that are appropriate to the situation and linguistic context. This capacity, which transcends stimulus–response associations ("stimulus independence"), has long been central to accounts of human linguistic capacities (e.g., Descartes). Early in his career, Chomsky envisioned an intimate connection between generative linguistics and CALU (e.g., Chomsky, 2011, p. 8, 1965, p. 6), especially insofar as recursive syntax provides a foundation for the unboundedness of linguistic competence. But he later came to endorse a mysterian stance regarding CALU as a whole: "Honesty forces us to admit that we are as far today as Descartes was three centuries ago from understanding just what enables a human to speak in a way that is innovative, free from stimulus control, and also appropriate and coherent" (Chomsky, 2006, p. 11). He added that CALU will likely always remain beyond the reach of scientific inquiry (Chomsky, 1982, 2006; see Drach, 1981, for further discussion).

LLMs pose a stark challenge to mysterianism. As human-built systems that generate fluent, contextually appropriate responses, LLMs instantiate many features of CALU. For example, LLMs exhibit humanlike levels of novelty (McCoy et al., 2023), pragmatic reasoning (Lipkin et al., 2023), and stylistic adaptation (Kandra et al., 2025), and humans cannot reliably distinguish human versus LLM conversation partners (Jones et al., 2025). Thus, LLMs are positioned to be the first tangible scientific models of key aspects of CALU, a promising step in moving beyond mysterianism toward mechanistic understanding.

Several LLM mechanisms are likely central to a future "science of CALU." First, LLMs operate with massive feature spaces, with recent work suggesting that small models (17 billion parameters) encode millions of features in dense superposition (Templeton, 2024), and that large models with trillions of parameters likely contain dramatically more features. The vastness of LLM representational regimes allows encoding of countless shades of meaning and extraordinarily fine-grained distinctions among contexts, enabling the system to retrieve and combine just those specific features relevant to producing situationally apt utterances. Second, attention mechanisms selectively amplify subtle aspects of the context, supporting fine-grained sensitivity to discourse and ongoing temporally evolving inter-speaker dynamics. Third, autoregressive generation factors predictive distributions over extended sequences into tractable components (involving predictions of next tokens conditional on context), enabling the open-ended production of never-before-seen utterances and the emergence of genuine novelty. Fourth, reinforcement learning supports policy learning from iterative feedback, aligning language generation with situational goals and pragmatic appropriateness. Together, these mechanisms constitute an initial toolkit for a future science of CALU, and the next generation of work will elaborate them in richer detail and add other mechanisms.

Importantly, LLMs also clarify why earlier attempts to build computational models of language generation and use – whether in cognitive science or artificial intelligence – were destined to fall short of an adequate account of CALU. LLMs suggest that CALU is an emergent property of *scale*: it arises only when an architecture has access to extraordinarily rich representational resources and correspondingly powerful mechanisms for manipulating them (Sutton, 2019). By contrast, prior approaches to modeling language use – e.g., assistive dialog systems or question-answering systems – were heavily engineered and tailored for very limited domains, and relied on correspondingly limited representational resources. The gap between even the best of these pre-LLM models and human-

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

level CALU was so large that Chomsky's mysterianism remained an appealing stance. LLMs embody a sharply contrasting approach: the ability to flexibly tailor linguistic production to nuanced contexts requires vast numbers of features and the capacity to deploy them through similarly vast numbers of tightly interlocked operations. LLMs provide, for the first time, a viable account of CALU as an emergent ability of high-dimensional, high-capacity models – a decisive shift in the explanatory paradigm.

To be sure, there are also critical divergences between human CALU and the analog observed in LLMs, but they need not be viewed as decisive impediments to the overall approach. Instead, these divergences represent promising research directions within a new framework. For example, LLMs often exhibit pragmatic brittleness, such as difficulty with irony, indirect speech acts, or norm-governed conversational implicatures (Hu et al., 2023). Each of these divergences highlights a concrete locus where mechanistic hypotheses can be sharpened, and understandings of new, heretofore poorly appreciated mechanisms might emerge. And of course, the most impressive capacities for CALU reside in models trained on internet-scale data. Rigorous poverty-of-stimulus arguments might one day motivate specialized innate mechanisms behind CALU in humans. But we are only in the earliest stages of understanding what is possible in building models with more limited data, and more importantly, what is possible with data derived from grounded, agentic experience (Vong et al., 2024).

One reason generative linguistics has historically focused narrowly on grammatical competence is that questions like CALU seemed intractable to formal analysis; indeed, Chomsky began sketching arguments along these lines as early as his 1959 review of Skinner's *Verbal Behavior*. But LLMs shift the boundary of tractability. They supply not only successful demonstrations of CALU-like behavior but also concrete mechanistic hypotheses for how such behavior might arise. By treating LLMs as artificial model systems, linguistics gains powerful tools to study the "inner springs" of creative language use at levels once dismissed as scientifically inaccessible.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing interests. None.

References

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press.
Chomsky, N. (1982). A note on the creative aspect of language use. *The Philosophical Review*, *91*(3), 423–434.
Chomsky, N. (2006). *Language and mind*. Cambridge University Press.
Chomsky, N. (2009). *Cartesian linguistics: A chapter in the history of rationalist thought*. Cambridge University Press.
Chomsky, N. (2011). *Current issues in linguistic theory* (Vol. *38*). Walter de Gruyter.
Drach, M. (1981). The creative aspect of Chomsky's use of the notion of creativity. *The Philosophical Review*, *90*(1), 44–65.
Hu, J., Floyd, S., Jouravlev, O., Fedorenko, E., & Gibson, E. (2023). A fine-grained comparison of pragmatic language understanding in humans and language models. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4194–4213.
Jones, C. R., Rathi, I., Taylor, S., & Bergen, B. K. (2025). People cannot distinguish GPT-4 from a human in a Turing test. *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 1615–1639.
Kandra, F., Demberg, V., & Koller, A. (2025). LLMs syntactically adapt their language use to their conversational partner. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 873–886.
Lipkin, B., Wong, L., Grand, G., & Tenenbaum, J. B. (2023). Evaluating statistical language models as pragmatic reasoners. *arXiv Preprint arXiv:2305.01020*.
McCoy, R. T., Smolensky, P., Linzen, T., Gao, J., & Celikyilmaz, A. (2023). How much do language models copy from their training data? Evaluating linguistic novelty in text generation using raven. *Transactions of the Association for Computational Linguistics*, *11*, 652–670.
Sutton, R. (2019). The bitter lesson. *Incomplete Ideas (Blog)*, *13*(1), 38.
Templeton, A. (2024). *Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet*. Anthropic.
Vong, W. K., Wang, W., Orhan, A. E., & Lake, B. M. (2024). Grounded language acquisition through the eyes and ears of a single child. *Science*, *383*(6682), 504–511.

Sociopolitical ramifications of language models make them worth worrying about

Alayo Tripp*

Linguistics, University of Florida, USA
atripp@ufl.edu

*Corresponding author.

doi:10.1017/S0140525X26104786, e222

Abstract

Regarding the utility of language models for linguistic research, Futrell and Mahowald advance a *crackpot realism*, wherein the concerns of a powerful elite are portrayed as "realistic" in a sense which is technocratic and detached from broader human consequences.

Futrell and Mahowald attempt to reconcile two extreme views on the utility of language models (LMs) for studying linguistic theories, human language learning, and processing. However, the presented viewpoints constitute a false dichotomy, relying on several assumptions which do not necessarily stand true and positioning investigations into the syntactic structure of well-resourced written languages as essential to linguistic knowledge. While there is no denying the influential status of generative linguistics, broadly stating that "linguists" should "love language models" does not do justice to the present diversity of objectives and approaches in the field.

Specifically, Futrell and Mahowald constitute a re-enforcement of narrow and exclusionary perspectives on the significance of syntactic structure to the nature of human language. Moreover, the titular reference to the film Dr. Strangelove (Kubrick, 1964) raises the specter of highly dangerous and destructive technology while lending an air of inevitability to discussion of its applications. Despite setting expectations that it will engage with conceptions of risk, Futrell and Mahowald overall neglect any such analysis. It

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

instead advances a crackpot realism, wherein the concerns of a powerful elite are portrayed as "realistic" in a sense which is technocratic and detached from broader human consequences (Mills, 2000).

Arguing that the reality of linguistic structure underlies the success of LMs, Futrell and Mahowald declare that "compressibility *is* structure." However, we must consider how fundamentally lossy this compression is. The LMs (LMs) under discussion rely on text, and the limited "linguistic layers" transparently encoded therein. Contrastively, children receive richly diverse audiovisual information and variation in phonetic and prosodic forms, grounded in a social world. The conflicting positions described in Futrell and Mahowald share an imagined connection between linguistic structure and function which casually excludes sociolinguistic variation and its functional correlates. The result is that both camps effectively champion a rarefied form of synthetic data, largely excluding sociolinguistic behavior from the conception of what language is. In the context of understanding human language acquisition, this is a notable oversight, as social and phonetic attunement to familiar varieties is acquired early (Kinzler, Dupoux, & Spelke, 2007) and may substantially guide processes such as word learning (Tripp, Feldman, & Idsardi, 2021).

The assumption that model input data belong to a single variety positions *standard language ideology* as compatible with both approaches to describing language, despite the centrality of social and pragmatic functions to evaluating LM success. The ideology is more broadly visible in the habitual conflation of language ability and intelligence found in discourse on LMs (Mitchell & Krakauer, 2023) and in the putative connection between linguistic ability and a "language of thought" (Berwick & Chomsky, 2016, p. 84). As language scientists, we must confront, not evade, this ideology and the way it animates both approaches.

Despite calls to diversify the languages under study, the role of English as a central source of insight is not critically addressed. Reliance on the most well-resourced language is a matter of convenience which undermines the scientific rigor of studies probing LM abilities. Most world languages do not have such ample training data, and in fact many lack an adequate number of living speakers to fruitfully compare LMs to the learning and processing of these languages (Liu, Richardson, Hatcher, & Prud'hommeaux, 2022). Consequently, indigenous and endangered languages are again pushed to the periphery of linguistic knowledge.

Likewise, universal sociolinguistic abilities such as the acquisition of multiple grammars, selection of speech registers, and interpretation of social identity through language behavior are all disregarded as outside the "core" linguistic function of incremental predictive processing, positioning the monolingual generative capacity of Internet-scale LMs as more critical to our understanding of language than any living speaker of the world's many endangered languages. This evaluation aligns with the profit motive behind LLM development.

Futrell and Mahowald associate danger with the neglect of LMs as a source of insight, insisting that we may disregard them "at our peril." LMs, as understood in this frame, effectively erase ways of languaging which are not readily captured in large repositories of text strings. Multimodal signals, admixtures of codes, and the use of sign languages are all ignored, implicitly treated as complexity which is justifiably lost in message "compression."

In this way, linguistic structure is presumed discoverable through probing systems designed specifically to produce prestige language forms. Further, these efforts are framed with urgency. While the engineering of improved language technology clearly stands to benefit those who control proprietary software systems, broader human consequences of reliance on LMs entail the continued exclusion and subordination of marginalized subjectivities. Language technology disproportionately discriminates against and misrepresents minoritized language users (Joshi, Santy, Budhiraja, Bali, & Choudhury, 2020). Foregrounding models of written standard English reproduces this existing dynamic, enshrining a specific variety as a uniquely useful and iconic representation of language form, function, and human intelligence.

For language science to equitably serve humankind, it is necessary to view both generativist and connectionist linguistic traditions with skepticism, opposing their focus on majoritized paradigms (e.g., monolingualism and literacy) at the expense of acknowledging the true diversity of natural human language behavior. This commentary seeks to contextualize LM utility within a broader discourse regarding answers to two questions: 1) who should be considered stakeholders in investigations of linguistic structure? and 2) what are our ethical obligations to those stakeholders? The reality of LMs is that a narrow sample of language varieties associated with globally dominant social groups have been treated as offering pressing insight into the human condition, while minoritized language systems are relegated to the role of confirmatory testbeds, if they are considered at all. Treating politically minoritized, disabled, racialized, and gendered persons as stakeholders in these investigations requires us to regard LMs which disregard and distort these subjectivities (Wang, Morgenstern, & Dickerson, 2024) as obviously inadequate and potentially dangerous.

In seeking a way forward which appropriately leverages LMs for linguistic insight, we must be explicit regarding the utility of written text and oppose the characterization of sociolinguistic diversity as harmlessly excluded from technocratic distillations of linguistic knowledge.

Acknowledgments. I thank Zoey Liu for her feedback on the initial proposal submission.

Financial support. This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Competing interests. The author declares none.

References

Berwick, R. C., & Chomsky, N. (2016). *Why only us: Language and evolution*. MIT press.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 6282–6293). Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.560

Kinzler, K. D., Dupoux, E., & Spelke, E. S. (2007). The native language of social cognition. *Proceedings of the National Academy of Sciences*, *104*(30), 12577–12580.

Kubrick, S. (1964). Dr. strangelove: Or how I learned to stop worrying and love the bomb. Columbia Pictures Corp. presents, Columbia TriStar Home Entertainment.

Liu, Z., Richardson, C., Hatcher, R., & Prud'hommeaux, E. (2022). Not always about you: Prioritizing community needs when developing endangered

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

language technology. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th annual meeting of the association for computational linguistics* (volume *1*: Long papers) (pp. 3933–3944). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.272

Mills, C. W. (2000). *The sociological imagination*. Oxford University Press.

Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, *120*(13), e2215907120.

Tripp, A., Feldman, N. H., & Idsardi, W. J. (2021). Social inference may guide early lexical learning. *Frontiers in Psychology*, *12*, 645247.

Wang, A., Morgenstern, J., & Dickerson, J. P. (2024). Large language models cannot replace human participants because they cannot portray identity groups. arXiv preprint arXiv:2402.01908.

LLMs as a platform for studying constraint interaction: Motivation and challenges

Ethan Gotlieb Wilcox[a]* and Elissa L. Newport[b]

[a]Department of Linguistics, Georgetown University, USA and [b]Neurology and Rehabilitation Medicine, Georgetown University Medical Center, USA
ethan.wilcox@georgetown.edu
elissa.newport@georgetown.edu

*Corresponding author.

doi:10.1017/S0140525X26104610, e223

Abstract

Large Language Models (LLMs) can serve as tools for understanding how probabilistic constraints interact during language acquisition. To motivate such use cases of LLMs, we discuss several examples from allied fields, including neurobiology and animal behavior, of how soft constraints shape learning and development in cognitive systems. We end by outlining four challenges that LLM cognitive modeling should address in the coming decade.

A promise of LLMs: Modeling probabilistic constraint interaction

Like Futrell and Mahowald (F&M), we see one of the major promises of LLMs as a new tool for modeling soft constraint interactions. By including different constraints into an LLM's architecture or training objective, one can simulate the causal effect of such constraints on the resulting model behavior and assess their effects individually or together. Crucially, because language models (LMs) can be trained on a human lifetime (or more) of linguistic input, one can run simulations at a level of ecological validity that was not available to previous generations of researchers.

In this response, we wish to deepen and extend the discussion in F&M by connecting it to allied fields. We also raise important phenomena from human language acquisition that we suggest should be included in testbeds for LLMs.

Why study constraint interaction? Looking beyond deep learning

F&M motivate the power and promise of graded constraint interaction by pointing to the literature on deep learning and Artificial Intelligence (AI). However, evidence in allied fields also supports the perspective that learning biases in biological systems can be gradual, probabilistic, and, in some cases, domain-general.

As one example, consider *imprinting*, the acquired attachment system in ducks and geese. Imprinting is commonly explained as resulting from the interaction of probabilistic biases or gradients of salience during learning (Bateson, 1979; Hess, 1973). As soon as they can locomote, ducklings will follow moving objects, which produces attachment to the followed object. The onset of the ability to walk, therefore, opens a critical period for imprinting. The more intense the motor behavior, the stronger the imprinting, with no imprinting if the animal is passively carried and very strong imprinting if the animal must climb over hurdles on the imprinting track (Hess, 1958). There are also gradients in the properties of attachment objects: a real duck or a sphere elicits stronger imprinting than other shapes. And there is a gradual decline in the strength of imprinting with hours after hatching. The end of the critical period is produced by the onset of fear of novel objects: ducklings begin to run away from a new object rather than follow it. In interaction in the natural world, these gradients produce strong and long-lasting attachment to duck-like moving objects that are followed within a short time after the development of locomotion. In this example, a fairly complex behavioral system has evolved successfully not by innately specifying the behavioral outcome but rather by the interaction of preferences and saliences that conspire with normal environmental input to achieve successful duck life.

Research on what begins and ends critical periods in many behavioral systems also shows that their neural underpinnings are an interacting set of gradient processes. Learning and synaptic sprouting begin when excitatory and inhibitory inputs to a network are in place; they end when myelin and the extracellular matrix develop, locking in the connections that are most functional. All of these processes are dynamic, not fixed – they are affected by the strength of the inputs they receive – so that "critical periods" can begin and end at a variety of times, depending on the environment (Hensch, 2005).

The natural world abounds with complex systems whose behavior is determined by interactions between graded, fuzzy constraints. We are optimistic that human language acquisition can ultimately find a similar constraint-based explanation and see ANN modeling as one important route towards realizing this goal.

Four challenges for LLM modeling

Below, we outline several challenges to which cognitive modeling can contribute and which we argue should be prioritized in the next decade.

Learning beyond the input: Most studies with LLMs train models on data that contain the full structural complexity of the human language to be acquired and show that the model makes the generalizations exhibited in the linguistic input. However, when their input does not contain full complexity or consistency, children modify language structures, introducing regularities and new structures not present in the original data (Austin et al., 2022; Singleton & Newport, 2004). Such behavior is often absent in LLMs and, indeed, absent in adult learners. How can architectural

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

constraints or constraints on learning mechanisms give rise to childlike regularization behavior?

Anything Goes? An important claim in the linguistics literature is that human learners will acquire only languages that observe the principles of natural languages (Bickerton, 1984; Chomsky, 1965, 1981). This has been shown in artificial language learning experiments (Culbertson & Newport 2015, 2017; Newport 2016) and for LLMs (Kalini et al., 2024). However, this preference appears to be weaker in LMs than in people (Yang et al., 2025). Children will dramatically change languages that violate linguistic universals to restore patterns common to human languages (Culbertson & Newport, 2015, 2017). What types of LLM mechanisms might reproduce these strong constraints?

Explaining Hierarchy: Our theory of hierarchy in language, roughly that people learn rules within a single context-free grammar, has not changed in 70 years. Yet LLMs obtain linguistic abilities without the computational capability to represent such grammars (Hahn, 2020). We must explain how and why soft constraints give rise to something equivalent to hierarchical LLMs, what this consists of, and whether it necessitates a new theory of hierarchy and structure in human language.

Human-Sized Data: While we are sympathetic to the authors' "Contravariance Principle," we believe that learning on 30 trillion words (LLMs) vs. 100 million words (roughly, humans) presents a vast difference in problem space, enough to question the contravariance principle. Linguists interested in computational modeling should run experiments on human-sized models (Warstadt & Bowman, 2022; Wilcox et al., 2025). Thus far, "BabyLM" models fall short of people, implying that current architectures are not sufficient to make the correct linguistic generalizations from a plausible amount of training data (Warstadt et al., 2023). How can soft constraints enable LLMs to learn from a biologically plausible data size?

Acknowledgements. None.

Financial support. Ethan Gotlieb Wilcox: Funding from Georgetown University and the Department of Linguistics. Elissa L. Newport: Funding from NSF grant BCS2445381, the George Bergeron Endowment, the Feldstein Veron Innovation Fund, and the Center for Brain Plasticity and Recovery at Georgetown University Medical Center.

Competing interests. None.

References

Austin, A. C., Schuler, K. D., Furlong, S, & Newport, E. L. (2022). Learning a language from inconsistent input: Regularization in child and adult learners. *Language Learning and Development*, *18*, 249–277.

Bateson, P. (1979). How do sensitive periods arise and what are they for? *Animal Behavior*, *27*, 470–486.

Bickerton, D. (1984). The language bioprogram hypothesis. *Behavioral and Brain Sciences 7*, 173–221.

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press.

Chomsky, N. (1981). *Lectures on government and binding*. Foris Publications.

Culbertson, J., & Newport, E. L. (2015). Harmonic biases in child learners: In support of language universals. *Cognition*, *139*, 71–82.

Culbertson, J., & Newport, E. L. (2017). Innovation of word order harmony across development. *Open Mind: Discoveries in Cognitive Science*, *2017*(1), 91–100.

Hahn, M. (2020). Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics*, *8*, 156–171.

Hensch, T. K. (2005). Critical period plasticity in local cortical circuits. *Nature Review Neuroscience*, *6*(11), 877–888.

Hess, E. H. (1958). Imprinting in animals. *Scientific American*, *198*, 81–90.

Hess, E. H. (1973). *Imprinting: Early experience and the developmental psychobiology of attachment*. Van Nostrand-Reinhold.

Kallini, J., Papadimitriou, I., Futrell, R., Mahowald, K., & Potts, C. (2024). Mission: Impossible language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers) (pp. 14691–14714).

Newport, E. L. (2016). Statistical language learning: Computational, maturational and linguistic constraints. *Language and Cognition*, *8*, 447–461.

Singleton, J. L., & Newport, E. L. (2004). When learners surpass their models: The acquisition of American Sign Language from inconsistent input. *Cognitive Psychology*, *49*, 370–407.

Warstadt, A., & Bowman, S. R. (2022). What artificial neural networks can tell us about human language acquisition. In S. Lappin, & J.-P. Bernardy (Eds.), *Algebraic structures in natural language* (pp. 17–60). CRC Press.

Warstadt, A., Mueller, A., Choshen, L., Wilcox, E., Zhuang, C., Ciro, J., Mosquera, R., Paranjape, B., Williams, A., Linzen, T., & Cotterell, R. (2023). Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM challenge at the 27th conference on computational natural language learning* (pp. 1–34).

Wilcox, E. G., Hu, M. Y., Mueller, A., Warstadt, A., Choshen, L., Zhuang, C., Williams, A., Cotterell, R., & Linzen, T. (2025). Bigger is not always better: The importance of human-scale language modeling for psycholinguistics. *Journal of Memory and Language*, *144*, 104650.

Yang, X., Aoyama, T., Yao, Y., & Wilcox, E. G. (2025). Anything goes? A crosslinguistic study of (im)possible language learning in LMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), (pp. 26058–26077).

Authors' Response

You can't fight in here! This is BBS!

Richard Futrell[a]* and Kyle Mahowald[b]*

[a]University of California Irvine, USA and [b]The University of Texas at Austin, USA
rfutrell@uci.edu
kyle@utexas.edu

*Corresponding author.

doi: 10.1017/S0140525X26105123, e224

Abstract

Norm, the formal theoretical linguist, and Claudette, the computational language scientist, have a lovely time discussing whether modern language models can inform important questions in the language sciences. Just as they are about to part ways until they meet again, 25 of their closest friends show up – from linguistics, neuroscience, cognitive science, psychology, philosophy, and computer science. We use this discussion to highlight what we see as some common underlying issues: the *String Statistics Strawman* (the mistaken idea that LMs can't be linguistically competent or interesting because they, like their Markov model predecessors, are statistical models that learn from strings) and the *As Good As It Gets Assumption* (the idea that LM research as it stands in 2026 is the limit of what it can tell us about linguistics). We clarify the role of LM-based work for scientific

insights into human language and advocate for a more expansive research program for the language sciences in the AI age, one that takes on the commentators' concerns in order to produce a better and more robust science of both human language and of LMs.

R1. Introduction

Our position is: language models do not replace linguistic theories, but they *do* tell us things about language, and they *do* force us to rethink certain approaches. Linguistics shouldn't become LLMology, but it should have a seat at the scientific table where these models are being produced and analyzed – contributing to that science and learning from it.

We see this as a middle ground position, and some commentators agree. Millière calls our position "a compelling middle path," and Capone and Lenci describe it as "more conciliatory and readily acceptable" than other work advocating for LMs as linguistic models. But others read our position as a disguise for extremism. Howitt and Lidz write that we "present ourselves as offering a measured middle ground" while aligning "aggressively" with those who think LLMs obviate linguistic theory. Rawski and Heinz suggest our "centrism is an illusion": "By the end, the sheep becomes a wolf: Claudette is inviting Norm for dinner, and Norm is on the menu."

In fact, we have no interest in eating Norm. We are serious that LMs don't replace linguistic theories. We believe that linguistic theory, as instantiated across a variety of formalisms, contains the best-known characterization of human linguistic competence. If you want an effective theory of how people judge the grammaticality of new sentences, how human languages vary, how native speakers project their limited experience to an infinite variety of new utterances, and how to capture these things in terms of mathematical structures, you will find it in the linguistics literature. We think that linguistic theory, as the result of decades of careful intellectual labor, has identified many of the most important real patterns of language – patterns that have been fruitful for understanding the behavior of both brains and language models. In our own work on LMs, we have drawn on ergativity and case-marking (Papadimitriou et al., 2021), the dative alternation (Yao et al., 2025), filler-gap dependencies (Wilcox, Futrell, & Levy, 2018, 2024), minimal pair grammaticality judgments (Hu et al., 2026), the learnability of languages with linear rules (Kallini et al., 2024), garden-path sentences (Futrell et al., 2019), idiosyncratic syntactic constructions (Misra & Mahowald, 2024), and many other linguistic phenomena. None of this work would have been possible without linguistics. If you want to understand how language works (in a brain or in an LM), in practice, you have to do so with reference to constructs from linguistic theory or else reinvent them.

The alternative is pessimism about whether complex cognition like language is explainable at all, and in fact, we find this view in the AI community and literature. On this view, the "right answer" about complex cognition is that it is too complex for a human-understandable theory, and mechanistic interpretation is a misguided quest; for language, this would mean that the black box is fundamentally opaque and interpretability is intractable. Geoff Hinton articulates a version of this stance: "If you put in an image [into a neural network], out comes the right decision, say, whether this was a pedestrian or not. But if you ask "Why did it think that" well if there were any simple rules for deciding whether an image contains a pedestrian or not, it would have been a solved problem ages ago" (quoted in Simonite, 2018). Hendrycks and Hiscott (2025): "It may be intractable to explain a terabyte-sized model succinctly enough for humans to grasp." The pessimism here runs deep: not just that AI is too complex, but that complex behavior is too complex to ever be amenable to explanation. This attitude is, in part, responsible for an estrangement between much of the community developing LMs on one hand and linguistics, cognitive science, and even NLP on the other.

We think linguistics, as a discipline and as a body of ideas, can help fight that pessimism. But if linguistics sits out because it treats present shortcomings as decisive – because models don't go meaningfully beyond string statistics, are insufficiently multilingual, learn patterns that are unlikely to show up in a human language, have a learning trajectory too different from that of humans, don't give insight into formal compositional meaning, are too black box for good science, or are too cognitively alien – then we think everyone loses. AI loses decades of conceptual tools for understanding complex minds. Linguistics loses contact with the only systems other than the human brain that demonstrate the phenomenon of normal language use.

We agree with some of the criticisms above regarding the limitations of LMs *if we were to take these models as replacements for linguistic theory* or *if we were stuck forever doing work only on closed-source, text-based models of English*. But we don't take them that way. We think LMs can serve as model systems: they show how linguistic structures can be learned and represented in a system that is rather unlike what has been assumed before, and in doing so they generate strong hypotheses for how the human brain does it, and they demonstrate ways of thinking about learning and representation that might be new to formal linguists, although perhaps less new to linguistics as a broader intellectual tradition. Neural networks have been indispensable in this role in other areas of cognitive science, like vision. Conversely, linguistics provides concepts that neural network interpretability researchers need to explain linguistic behavior in LMs and to enhance it.

In response to current limitations, we advocate an expansive approach that challenges these limitations head-on. Make work with LMs more multilingual and inclusive (Tripp, Ármannsson et al.). Bring audio into the fold (de Heer Kloots et al.). Run studies focused on data size and diversity (Ambridge, Wilcox, and Newport; Lampinen) and the learning trajectory (McDermott-Hinman and Feiman; Howitt and Lidz). Do projects involving evolution, communication, and interaction (Culbertson et al.; Bunzeck et al.; Sripada et al.). Explore what kinds of languages LMs can and can't learn well and what that tells us about inductive bias (Bowers and Mitchell; Kallini and Potts; Rawski and Heinz). Study how models got so good at language (Lupyan) and where they still have weaknesses (Murphy et al.; Bolhuis et al.), doing work in mechanistic interpretability to find out which representations are real (Nefdt and Ladyman; Millière) and if they instantiate generative linguistic structures (Goodale and Mascarenhas; McCoy) or distributional ones (Capone and Lenci). Develop more rigorous theory around how they map onto human cognition and human language – and what we can learn from that mapping (Resnik; Nair and Phillips). Make the work reproducible, rather than a black box and opaque (Dingemanse and Cuskley).

Our bet is that all of these avenues are promising. And that's why we see ourselves as carving out a space between those who just want to keep building models without taking linguistics seriously and, on the other hand, those who think the models are an artifact best to be ignored in serious language science. Having said that, as we make clear both in our target article and in this response, we are not agnostic about the role that LMs can play in language science. Here, we lay out the case that they will continue to be revolutionary

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

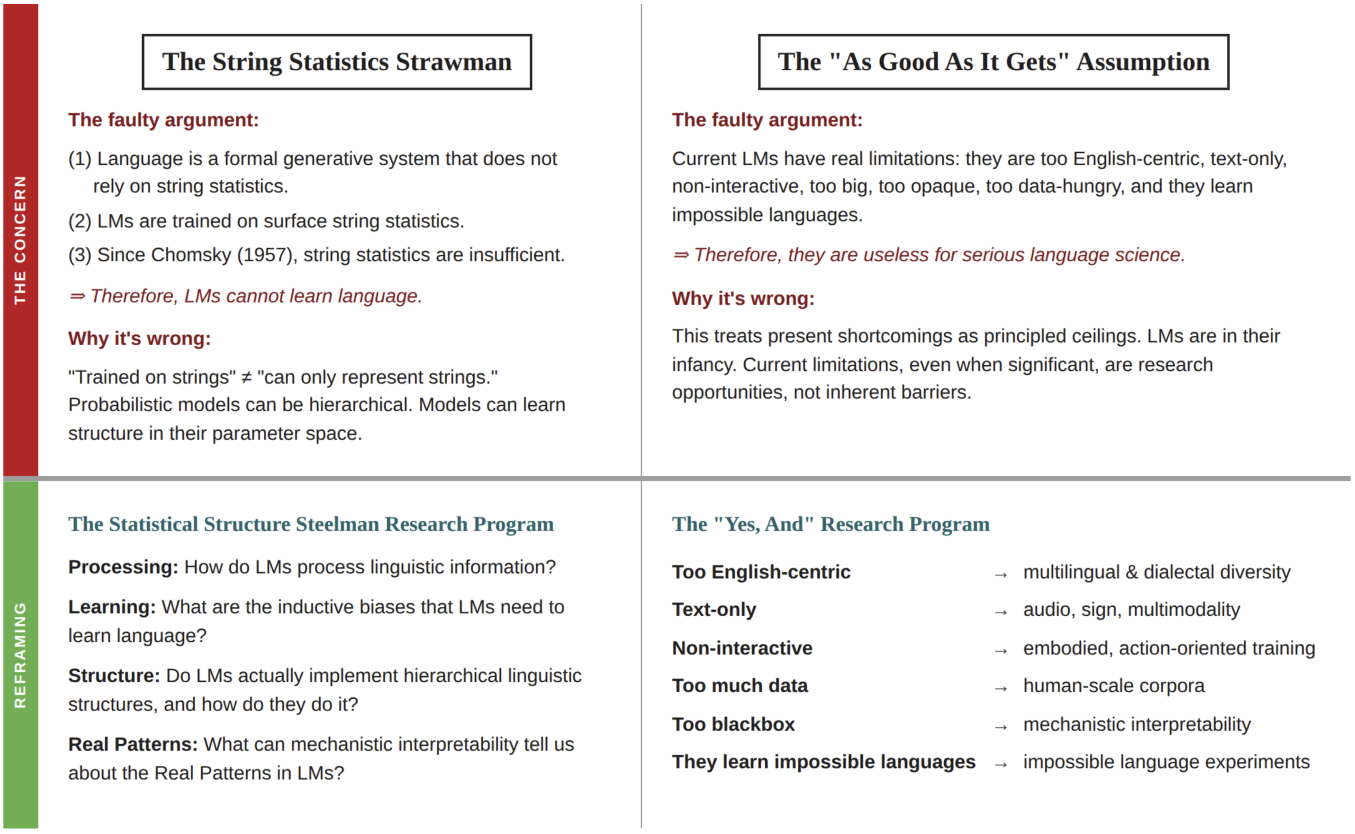


Figure 1. Two common fallacies about LMs and linguistics.

for scientific inquiry into human language – especially when studied in light of linguistic theory and especially as some of their present limitations get lifted.

First, we diagnose what we take to be a central mistaken premise in many objections to LMs in linguistics: the *String Statistics Strawman* (see Figure 1). The String Statistics Strawman roughly states that language is richly structured and not string-based, that statistical string-based systems have long been known to be insufficient for learning language, and, therefore, LMs either don't really learn language or do so in a way that's irrelevant for interesting linguistic questions. We argue that, contra the String Statistics Strawman, the nature of LMs is perfectly compatible with their representing rich, hierarchical, and generative linguistic structure. We argue that they really do learn language and reject claims that they in principle cannot have linguistic competence.

Second, we address a set of skeptical responses that don't draw upon the String Statistics Strawman but that still doubt how much we can treat LMs as model systems. We clarify what we think we can learn theoretically from LMs about human cognition (which we think is a lot!), with a focus on processing, learning, and structure. In particular, we explore what LMs reveal about inductive bias in language and clarify whether they are compatible with generative linguistics (they are) or are a success for usage-based distributional linguistics (they also are). In doing so, we develop our account of linguistic structures as real patterns.

Third, we received a number of commentaries highlighting shortcomings of LM-based linguistics research as it currently exists. We agree with many of these commentaries and lay out a vision for a positive research program, highlighting that many of the *objections* to LM-involved language science can be reframed as productive research ideas. At the same time, we caution against the *As Good As It Gets Assumption* (see Figure 1), assuming that the current limitations of LM research (too much English, models too big and too closed-source, not enough interaction) are *in principle* limitations.

Finally, we conclude with a set of broader thoughts responding to the current atmosphere of fear and desire around modern AI, emphasizing that we can embrace language science while remaining clear-eyed about the advantages and disadvantages of the technology.

R2. Against the String Statistics Strawman

Many of the skeptical responses to LMs in the literature have a similar flavor that we think reflects a common origin. Call it the String Statistics Strawman:

(1) Language is a formal generative system that generates grammatical sentences based on structural representations.
(2) LMs, like Markov models, are statistical models trained on surface string statistics.
(3) Since Chomsky (1957), it is well established that surface string statistics are not sufficient to characterize language in the sense of (1).
(4) Therefore, LMs do not model language in the sense of (1).

We can trace this line of thinking to the canonical demonstration in Chomsky (1957) – "Colorless green ideas sleep furiously" – and the observation that simple Markov models cannot distinguish grammaticality from corpus frequency. But the

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

capacities and generalization behaviors of modern LMs are fundamentally different from those of early *n*-gram language models; in particular, they can go beyond string statistics. Invoking early failures of simplistic Markov models to delimit what modern LMs can, in principle, represent is like invoking early Wright Brothers flights to argue that jets cannot cross the Atlantic. Thus, the step from (2) and (3) to (4) simply does not go through. The inference from "trained on strings" to "can only represent strings" is not right for LMs, akin to concluding that, because humans learn language from acoustic signals, they can only represent acoustic patterns.

Bolhuis et al. illustrate this view: LMs are "useless for linguistics" because they are "probabilistic models" that require a "vast amount of data" to analyze "externalized strings of words," whereas human language is "underpinned by a mind-internal computational system that recursively generates hierarchical thought structures." Everything in that description can be true, and the conclusion can still be false. LMs are trained on data (sometimes vast amounts, sometimes less). They are probabilistic models (although the brain is also probabilistic). And they are trained on externalized strings (although human learners also are faced with a continuous stream of input and not simply handed parsed hierarchical data structures). But it doesn't follow that this means they can't learn an internal system that generates hierarchical thought structures. Probabilistic models can be hierarchical – probabilistic context-free grammars are both; even minimalist grammars can be probabilistic without much trouble (Hunter & Dyer, 2013). Systems trained on strings can learn latent hierarchical structure, as shown by the large literature on grammar induction (Clark & Lappin, 2010b). There is no tension between outputting string probabilities and being recursive and hierarchical.

In fact, neural sequence architectures can in principle implement rich generative systems, abstracting far away from surface statistics (Allen-Zhu & Li, 2025; Bhattamishra, Ahuja, & Goyal, 2020; Cagnetta & Wyart, 2024; Cagnetta et al., 2024; Hewitt et al., 2020; Korsky & Berwick, 2019; Lan, Chemla, & Katzir, 2022, 2024a; Weiss, Goldberg, & Yahav, 2018). Whether they *actually do* implement such systems in any given case is empirical, but dismissing the possibility on the grounds that they have a "probabilistic nature . . . in complete contrast to the recursive function by which the human mind generatively forms hierarchical thought structures" (Bolhuis et al.) can only make sense under the String Statistics Strawman.

Once we acknowledge that LMs *can* represent abstract structure, we open up the whole landscape of ongoing empirical and theoretical work about what they actually do. The big questions are: First, do LMs actually learn the kind of structure linguists care about? Second, even if they do learn it, is their competence (whatever it is) of *linguistic interest* – or is it too alien, too uninterpretable, too divorced from meaning, learning, and use? The next sections take these questions in turn.

R3. Do LMs learn language (in the relevant sense)?

It is now possible to have long, fluent, coherent conversations with frontier LMs. It is hard to maintain the view that these frontier models do not understand language at all – even if one thinks they do so in a way that differs from humans, even if one thinks that models trained on human-scale data would behave differently, and even if the nature of that understanding may change as models and methods evolve. We agree with Lupyan (contra Murphy et al. and Bolhuis et al.) that "large language models have learned to use language" and that these systems "actually work." We think large models have learned enough linguistic structure to be interesting to linguists and (self-evidently) enough to demonstrate mastery in normal usage.

Today's language models do not rely purely on shallow heuristics. In many cases, earlier models *were* relying on heuristics that could lead to the wrong behavior. For example, McCoy, Frank, and Linzen (2018) show that recurrent networks trained to transform declaratives into questions end up learning a linear heuristic rather than a hierarchical generalization. But Ahuja et al. (2025) show how, on the same training dataset, a different (and more realistic) training objective does guide LMs to the desired hierarchical generalization in a way that reflects principles of Bayesian learning from indirect evidence (Perfors, Tenenbaum, & Regier, 2013). LMs do represent hierarchical structure – the *sine qua non* of language under some views – as revealed not only by extensive and careful behavioral tests but also, more importantly, by probing and causal interventions that begin to answer *why* the model outputs what it does (Geiger et al., 2025; Mueller et al., 2026). The representations revealed through these methods are abstract, generalizing across different languages (Brinkmann et al., 2025; Papadimitriou et al., 2021). It isn't just shallow string statistics or wrong heuristics.

At the same time, we don't think the LMs have matched or done better than linguistic theory at giving explanatory descriptions of human language. Their latent parses can be wrong, and the representations do not always generalize as we think humans do. Nevertheless, we are not willing to say that they have no linguistic structure simply because their representations do not look exactly like a particular theory of syntax or because they do not parse every sentence as a human would. They represent the *kinds* of structures posited by linguists, and they do so by embedding them in an interesting and unexpected way in a continuous vector geometry.

Murphy et al., in contrast, write: "LLMs can surely capture certain statistical regularities from the output of language (text, data), but they do not infer grammar the way humans do and fail to capture fundamental principles of linguistic structure [(Murphy et al., 2025)]." The cited paper (Murphy et al., 2025) holds models to what we think is not a reasonable standard: in one prompt (admittedly cherrypicked here; others are less fanciful; see Lupyan for further discussion and critiques of this set of experiments), frontier models are prompted with "Pretend that 'glart' is a word that refers to a group of alien creatures, and can also refer to the action of pleasing. In this context, is 'Glarts glarts glart glart glarts glarts' grammatical?" Even as professional linguists, we aren't sure how to answer this question: should we assume that any lexical item that can express "the action of pleasing" can be realized as a verb? If so, is it transitive or intransitive, and does it take the pleasee as the subject or the pleaser? Nonetheless, we tried this prompt with the latest versions of ChatGPT and Claude, and what they outputted was linguistically excellent.[1]

Can this really be a test that a model must pass in order to give insight about language? Whether models can do this kind of metalinguistic task is interesting in its own right, but it's interesting as an AI-for-science challenge. We don't see this kind of metalinguistic task as a prerequisite for whether models have learned grammar (Hu et al., 2024); many humans would be confused by the question and would not do well either. Furthermore, the LM might be deploying an entirely different system when answering explicit metalinguistic questions as compared to when generating normal text. Some of the objections seem to concern the use of experimental methods, often derived from psycholinguistics, to assess and probe these abstract representations in LMs. Rawski and Heinz think we "want linguistics to effectively become a niche within psychology (Núñez et al., 2019), where replicability and theory crises (Eronen &

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

Bringmann, 2021) mean 'theories rise and decline, come and go, more as a function of baffled boredom than anything else; and the enterprise shows a disturbing absence of that cumulative character that is so impressive' (Meehl, 1978)." We're somewhat baffled about how what we wrote could have led to this idea. But this sneering dismissal of an entire productive field of inquiry (psychology) is misplaced and exemplifies worrisome linguistic isolationism. For experimental studies of LMs in particular, the empirical work shows a strong cumulative character, with initial results leading to challenges and refinements, and with new results, methods, open datasets, and ideas building on old ones (e.g., the line of work on filler–gap dependencies: Boguraev, Potts, & Mahowald, 2025; Chaves, 2020; Gauthier et al., 2020; Kobzeva et al., 2025; Lan, Chemla, & Katzir, 2024b; Wilcox et al., 2018; Wilcox et al., 2024; *i.a*). Thanks to these methods and others, compared to 2018, we now have vastly more knowledge about how and when neural systems of different kinds implement linguistic structure.

R4. Is what they learn interesting for linguistics?

Even if LMs do learn and represent nontrivial linguistic structure, then it is possible that they remain uninteresting for linguistics – inasmuch as linguistics is about language as a product of human minds – because the structures they have learned, and the way they have learned them, are irreducibly alien compared to whatever humans have. We take this objection in pieces, focusing on the processing, learning, and structure of language.

R4.1. Processing: Sign us up for cognitive xenobiology!

Resnik and Nair and Phillips offer measured skepticism about what large language models can tell us about human language processing. They do not deny that LMs exhibit striking linguistic behavior, nor do they dismiss them as mere string-based heuristics. Rather, their concern is about the kinds of explanations LMs can legitimately support. Drawing on Marr's levels of analysis (Marr, 1982) in both cases, these commentaries emphasize that matching human behavior at the computational level does not by itself license conclusions about the algorithmic or implementational levels that underlie human sentence processing.

Resnik doubts the usefulness of LMs as cognitive models at all three of Marr's levels. He says the implementation is so obviously different as to not be useful. There are indeed important differences, and we certainly agree that figuring out the right way to abstract away from implementation differences between models and humans is important work for the kind of program we lay out. Yet results from vision show striking parallels between, for example, neural networks and the visual system, as we discussed in our target article, and different architectures can converge to similar representational ideas, which then form good hypotheses for brain representations (Hosseini et al., 2024; Huh et al., 2024). Furthermore, recent results with LMs show intriguing ways that details of the transformer architecture are more closely related than one might have expected to models of working memory in sentence processing (Ryu & Lewis, 2025). So, while we agree with Resnik's claim that modern AI engineering has moved away from its biologically inspired connectionist roots, we think the new implementations can still be informative.

At the computational level, Resnik seems to acknowledge that there may be behavioral parallels between models and people. But that by itself is, he says, unsatisfying because it might not be linkable to the algorithmic level: "There remains the burden of mapping any such computational theory to a biologically plausible algorithmic account of real-time processes distributed over diverse, specialized components." So we take his most pressing concern to be about the algorithmic level. At the algorithmic level, he particularly worries about mechanistic exceptions that would undermine our claim that we should expect convergence to similar mechanisms. Nair and Phillips share this worry about mapping potentially powerful computational-level parallels onto algorithmic and implementational ones. They see LMs as (usefully) supercharging the project of using predictive measures to estimate linguistic difficulty and complexity. But, like Resnik, they see that project as distinct from the psycholinguists' goal of "characterizing the underlying mental computations from each step of incremental processing."

Resnik raises a useful thought experiment: imagine aliens come to Earth, learn to produce language by reading all the text ever produced, and we can probe their brains. He asks: "[W]hy in the world would one expect to find mechanistic correspondences between what's happening in [the aliens'] boxes and what's happening in the human cognitive apparatus?" We agree that the thought experiment is a good one for our current situation. But we don't share the intuition that the alien brain would be so different from ours that it would have nothing to teach us. Partly, it seems that we have a different level of confidence in the expectation of convergence. Resnik doubts the contravariance principle in part because "[U]nlike visual object recognition, a key part of what makes "language" hard is that it decomposes into very distinct subproblems." In contrast, we think the parallel to vision is apt and that the focus on breaking language down into sub-problems (e.g., first sound-level, then syntax, then semantics, then pragmatics) has been overemphasized in psycholinguistics, with much evidence pointing towards a much more unified processing of these subproblems (MacDonald et al., 1994; Spivey & Tanenhaus, 1998; Tanenhaus et al., 1995; Trueswell, Tanenhaus, & Garnsey, 1994).

Of course, just because convergence is possible in theory doesn't mean that it really does happen. This caution motivates us to want to do more work in this domain, as we think both Resnik and Nair and Phillips would endorse, and proceed with scientific rigor. And we agree that one must be explicit about which level of explanation one is targeting and what evidence would bear on it. But, partly, it seems that we have a different reading of the last few years of linguistics-focused LM research. We think there is already a growing and important body of work (summarized in our target article) finding interesting parallels between the kinds of representations posited in linguistics and the kinds of representations emergent in neural models. As we discuss in Section R4.3.3, there is philosophical and cognitive work to do to learn how to think about the mappings across these very different systems. But we think a picture is emerging.

As with any scientific project, there are likely to be unexpected findings and developments along the way. But we find it hard to imagine a scenario where probing the linguistically competent aliens' brains didn't give us insight, one way or another, into human cognition. That can be true even though there will surely be divergences. Resnik cites as a cautionary tale the divergent mechanisms for prey capture in sharks (electroreception) and dolphins (echolocation). But, if there were dolphin scientists studying prey capture, and suddenly a bunch of sharks with highly probeable brains landed in the dolphin scientists' laps, we think they would learn a lot by studying the sharks and modeling how a similar computational goal is realized through a different mechanism. It would force them to interrogate their assumptions about what is required for prey capture, how different mechanisms

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

can work, whether there is a level of abstraction at which the dolphin mechanism can be mapped onto the shark one, and whether the different mechanisms lead to different prey-capture behavior. The discovery of sharks would be a great opportunity for the dolphin scientists, although we would recommend they proceed with care.

In short, now that the aliens are here, we think it would be scientifically negligent if no team of scientists tried to probe their language ability – and futile if that team did not involve linguists. So, sign us up for cognitive xenobiology!

R4.2. Learning: Reshaping, not denying, the learning problem

The responses also give us a chance to clarify our stance on innateness and inductive bias. Neural LMs succeed in learning language in ways that no previous system has, and so – even if the trajectory of their learning does not match humans' (McDermott-Hinman and Feiman, Howitt and Lidz) – understanding how they learn is relevant for understanding how language learning is possible. We wrote: "Far from demoting inductive bias as a concern, language models open up the range of (possibly innate) inductive biases that we might look for in humans. The main point for linguistic theory is not to demote the importance of the learning problem, but to reshape it: away from explanation by constrained description, and toward a broader landscape of approaches and hypotheses."

The central point is not that LMs argue against innate biases, but rather that those biases can look very different from what much of the generative linguistics literature assumed they had to look like. Inductive biases in large neural networks are less like an elegant discrete grammar formalism where the learner just needs to flip a few categorical switches and the rest follows deductively (Berwick et al., 2011; Chomsky, 1981; Gibson & Wexler, 1994). They are more like a collection of soft constraints that arise from a variety of sources. Some of these sources are domain-specific to language, and some of them are domain-general constraints that are useful for a variety of tasks. At the computational level, the inductive biases in large neural networks look less like a hard delimitation of possible hypotheses and more like the kinds of soft Bayesian priors that have appeared in computational modeling work on language acquisition (Hsu, Chater, & Vitányi, 2011; Perfors et al., 2013) – which may well be innate and language-specific (Pearl & Sprouse, 2013)!

Rawski and Heinz doubt that this idea of soft versus hard constraints in learning makes sense – indeed they question the idea that constraints can be domain-general or domain-specific at all. They point out, for example, that any learner operates under the hard constraint that it has a fixed computable hypothesis class and that the inductive biases we surveyed are specific to certain neural architectures, making them putatively domain-specific. Yet their discussion seems to use these terms in nonstandard ways that make them trivial. Indeed, neural learners cannot entertain uncomputable hypotheses – in this sense, all possible learners trivially have a hard constraint. But what is interesting about learning in large neural networks, and what was surprising even to statistical learning theorists (Zhang et al., 2021), is that the hard limits of their hypothesis space are less important for determining their inductive biases than the soft constraints that are imposed inside that space. This was a nontrivial discovery about how learning can work – although it has since been understood retroactively in terms of existing theories (Wilson, 2025) – and it should cause us to broaden our view of how human learning can work as well. Similarly, neural LMs clearly have biases that depend on their architectures, but nonetheless, they are domain-general learners in the standard sense of the term from cognitive science: their biases do not limit them to learning language (Clark & Lappin, 2010a; Pearl & Sprouse, 2013; Wilcox et al., 2024). If we take "domain-specific" to mean dependent on architecture as Rawski and Heinz seem to mean it, then all learners are trivially domain specific. This way of framing things does not seem productive.

Indeed, the domain-generality of neural LMs forms another objection to the idea that they can be useful for how we think about learning. Some would have it that if an LM could learn an "impossible language" – one violating constraints like a bias against grammar rules based on linear order – just as well as an attested human language, it would suggest LMs are so deeply unhuman-like that they are uninformative about human language. Bowers and Mitchell, Bolhuis et al., Murphy et al., Wilcox and Newport, and Kallini and Potts address this view from different perspectives.

Bowers and Mitchell claim that Kallini et al. (2024) found only two impossible languages harder than English and concluded no linguistic priors are needed to account for LM learning. This mischaracterizes both the findings and the claims in Kallini et al. (2024): they found that all tested impossible variants were harder to learn than English and explicitly attributed the gap to linguistic inductive biases – that is, inductive biases that are aligned with the structure of language, which may be either specific to language or general to many different tasks. Similarly, we claimed in the target article that LMs have inductive biases that are aligned with the structure of language and that discovering what these biases are is a fruitful endeavor for linguistics.

None of this implies that current LMs share *all* inductive biases with humans or that they can only acquire human language. Nor is that a requirement for them to be informative for linguistics. The inductive biases we discuss in the target article (information locality and low sensitivity) – as well as other inductive biases in neural LMs – seem to match real properties of language that were either not known before or not emphasized and that enable learning. Contra Bowers and Mitchell, Murphy et al., and Bolhuis et al., learning only possible languages is not a prerequisite for a model and its inductive biases to be informative. It is not even clear that humans have this property; subjects in Musso et al. (2003) learned the impossible variant of Italian to a similar level of accuracy and at a similar rate as the real language. There is no tension between being a domain-general learner and having biases that align with language: the biases that make LMs good at language, such as information locality, may be useful for many other tasks.

On the "impossible languages" question, we favor the characterization in Kallini and Potts (disclosing that we are coauthors with them on Kallini et al. (2024), a key study in this debate): the productive question is not whether LMs can or cannot learn impossible languages but what the pattern of their successes and failures reveals about which inductive biases support language learning. This reframing turns the impossible-language debate from a litmus test into a research program.

The upshot for theories of learning is that LMs do not argue against innate biases, but they argue against the view that the biases required for language learning must be narrowly linguistic and categorically hard. The success of domain-general learners at acquiring core properties of language tells us that the prior overgrammars can be softer – and broader in origin – than

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

previously assumed. Understanding exactly what that prior looks like is, we think, one of the most exciting questions that LM research opens up for linguistics.

R4.3. Structure: Generative AI isn't anti-generative – but it doesn't do generative linguistics any favors

R4.3.1. Whatever happened to structure? LMs and generative syntax

We agree with McCoy that it is entirely possible that LMs *could* instantiate theories based on formal structures, like those posited in generative linguistics. As such, we find McCoy's response an important antidote to the String Statistics Strawman, which assumes that nothing like LMs could ever learn language because it just isn't the right kind of system for learning language. But we also make the case that LMs are a success for distributional, usage-based accounts of language since they are, in the words of Capone and Lenci, "the most advanced product of distributional semantics." We do not consider these positions contradictory.

The reason we see LMs as a success for distributional approaches is that they solve what was long taken to be a fundamental limitation of distributional models: their inability to do rich, hierarchical composition (see Baroni, Bernardi, & Zamparelli, 2014, for an attempt to reconcile this). Despite being distributional learners, LMs clearly have some evidence of hierarchy and compositionality. As McCoy suggests, it's possible that they overcome their limitations by learning (or even being constructed so as to be predisposed to) certain formal structures. Or it's possible they do it some different way.

If it turned out that they learned formal generative machinery and/or benefited from it in training, we would find that a striking victory for generativism by showing that a minimally biased learner converged on the same representations posited by linguists. McCoy points to some intriguing evidence in that direction. But, unlike the case with distributional learners, nothing about LMs' success *as such* is a success for the generative framework. As we argue in the target article, the tradition they come from is one that is decidedly distinct from generative linguistics. A distributional semanticist or connectionist from the 1980s could be justified in looking at modern LMs and saying, "See, I knew it could work!" It's harder to make the case that a generativist could do the same – whereas they might have if these models were based on a scaling up of, say, the rich (and still important and interesting) tradition in minimalist parsing. This is not proof that one theory is right and the other wrong, but, if we were to update in one direction, it would be toward giving greater credence to the distributional semanticist/connectionist.

So, yes, LMs could be succeeding because they have learned generative linguistics. Testing that hypothesis is exciting, and we endorse McCoy's program for doing more of that work. But we see it as a claim to be tested and further explored. In contrast, LMs have *already* shown that usage-based, distributional approaches can be far richer and more capable than anyone realized.

R4.3.2. Whatever happened to meaning? LMs and formal semantics

Imagine a researcher interested in studying biological inheritance. Hugely important progress towards this goal was made by the Mendelian framework, with its elegant mathematical framework for showing how traits are passed down across generations. Since then, the understanding of biological inheritance has been dramatically improved by understanding the structure of DNA, studying large-scale population genetics with statistics, running experiments that causally intervene on genes in model organisms, and collecting massive data sets. None of that was possible in Mendel's time, but none of it obviates the important insights of the Mendelian framework. Rather, these developments help us better understand the framework. That said, it would be odd to become so focused on the mathematical framework that Mendelian probabilities become the end in and of themselves. That is, it would be odd to insist that to "really study biological inheritance," you have to bracket all the insights that come from a richer understanding of the implementation level and from statistics and computational modeling and just focus on the Mendelian patterns because all of the rest somehow "isn't really studying inheritance." We caution against going down a similar road when it comes to linguistics.

Goodale and Mascarenhas criticize the usage-based focus of our target article and what they see as a dismissal of generative linguistics, particularly work in "mathematically sophisticated compositional semantics pursued in mainstream generative-linguistics departments. In short, what they're doing over at MIT." They argue that our discussion of meaning mischaracterizes this strain of semantics by treating it as a competitor to more usage-based approaches. In their view, formal semantics "has never had much to say about the meanings of content words or the concepts that lie behind them" (although surely not all formal semanticists would agree with this, given the large volume of work in the formal semantic tradition on exactly these issues). For them, this omission is not a flaw but a methodological virtue. By abstracting away from content words and treating their meanings as placeholders, formal semantics can instead focus on the combinatorial principles that govern meaning composition.

We agree that formal semantics has been successful in developing a mathematically consistent framework, with genuine explanatory successes for quantifiers, conditionals, and a variety of other phenomena (including, in many cases, content words!). But it was developed with abstractions justified by limited explanatory ambitions about linguistic competence. LMs now give us ways to broaden those ambitions.

So we return to the commentators' question: whatever happened to meaning? If formal semantic structure is at the core of linguistic competence, it should be recoverable in systems exhibiting such competence. Without that connection, formal semantics would be a potentially interesting formal system but not a theory of language.[2] Retreating from explaining natural language in all its richness while focusing on the formal system – bracketing the meaning of *lion* or insisting we not "break faith with the mathematics" – remains an interesting project, but one that we think can now be expanded in dramatic and exciting ways. Formal composition is one important level of abstraction in a multi-level stack.

On this point, we refer to a body of work in the early 2010s (Baroni et al., 2014; Coecke, Sadrzadeh, & Clark, 2010; Socher et al., 2013), attempting to unify the seemingly divergent worlds of Montague semantics and distributional semantics by building fully vector-based realizations of formal semantic systems. Baroni et al.'s "Frege in Space," for instance, attempted to unify formal Montague semantics and distributional semantics by treating some expressions as vectors and others as operators over those vectors, composing meaning in a formally explicit way. The end result, if it all worked, would have been a vector-based system that implemented a full formal tree-like Montague semantics.

These hand-crafted approaches weren't tractable at scale. But, in principle, it is possible that modern LMs may emergently realize the "Frege in Space" vision: a fully worked-out vector-based

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

implementation of Montague semantics. Exploring this possibility and understanding the relationship between distributional representations and formal semantic abstractions is a promising research program, one that would be missed by insisting that the only way to study meaning is by preserving the mathematical purity of the compositional system.

One can tell a sympathetic historical story about why formal semantics developed in this way. Content words proved harder to formalize than logical operators (as Goodale and Mascarenhas note); integrating lexical meaning, world knowledge, and context was extraordinarily difficult; end-to-end semantic systems were invariably brittle. So the field focused where it could be rigorous. But LMs change what is tractable by enabling direct engagement with content words and context. We suspect the kind of formal semantics done at MIT will continue to bear fruit in the form of formal machinery for understanding compositional meaning – especially for the kind of questions that are difficult to study in systems like LMs, such as those that crucially depend on differences between literal meaning and use. But the new tools bring new opportunities, outside the old walls of tractability, to make progress on the very goals originally core to semantics.

R4.3.3. Real patterns meet mechanistic interpretability

So what, then, is the relationship between elegant formal theory and messy implemented systems? We argued that Dennett's (1989) concept of real patterns gives us a way to think about linguistic structure as it exists both in human minds and in neural networks. On this view, linguistic theory provides an effective theory of important aspects of observable linguistic behavior – it is valid and real on those terms, even when its relationship to the underlying reductive mechanisms is not clear. That is, linguistic theory remains real even if it does not correspond to a hard-coded innate circuit in a brain or in a neural net.

Two commentaries take up these questions, in particular. Nefdt and Ladyman question whether real patterns reside in the language or in the head and whether we got the real patterns account a bit backwards. Drawing on the famous Conway Game of Life analogy, Nefdt and Ladyman point out that in the Game of Life, complex patterns (the effective theory) emerge from simple deterministic rules (the reductive theory). But the way we see it, linguistic reality (whether in human brains, language models, or as an abstract system) is messy, complex, and fuzzy – and linguistic theory is a high-level account on top of that.

We don't see our view as contrary to real patterns. The Game of Life is underlyingly simple at the reductive level, but other canonical real patterns, like beliefs and desires, are notoriously complex at the reductive level. We think beliefs are a better analogy than the Game of Life: there is likely nothing strictly "in the head" that consistently corresponds to a belief – Dennett points out we believe things we've never thought, like "the moon is not made of cheese." Nonetheless, the concept of "belief" is fruitful for prediction. The point is that a pattern can be real – genuinely explanatory, not just a convenient fiction – even when it does not map neatly onto any single mechanism in the system that gives rise to it. We think the formal objects discovered by linguistic theory plausibly work in this way.

These questions raise important topics in mechanistic interpretability, the field that seeks to understand how AI models work mechanistically. Millière is sympathetic to the real patterns account but suggests we may actually *undersell* the potential contribution of mechanistic interpretability for understanding exactly how a linguistic task is solved. Our c-command example was intended to illustrate a contrast between what we might hope to find from these methods and what we actually find. The most promising recent methods rely on causal intervention to identify meaningful processes in LMs (Geiger et al., 2025; Mueller et al., 2026). As linguists or cognitive scientists, we ideally want to know: how is the LM performing complex linguistic tasks? Is it using classical linguistic representations and machinery to represent things like filler–gap dependencies, or has it found a different way to do it?

Because of their distributed nature, it's unlikely we are ever going to see something that looks so neat and tidy that it clearly identifies itself as a classical linguistic algorithm. Instead, we are able to make claims like, "By causally intervening on this particular part of the network, we can observe shared structure between related constructions" (Boguraev et al., 2025). But these claims require care. It's possible, for instance, to use an overly expressive interpretability model to find structure that isn't really there (Geiger, Harding, & Icard, 2025; Sutter et al., 2025). Nothing is going to replace good, careful science.

But that's not to say that things are hopeless: after all, brains aren't going to reveal neat and tidy algorithms either. Brains are distributed systems, and many of the computations they perform are approximate. If your criterion for "has linguistic competence" is "has a circuit that clearly, losslessly, and exactly implements a certain function," then you won't find it in a brain or in a giant neural net. But the task of finding structure in LMs using mechanistic techniques is a useful guide for understanding the relationship between lower-level processes and higher-level theory. Therefore, like Millière, we are excited about what mechanistic interpretability will do for questions in linguistics and cognitive science and the ability to "test, in unprecedented detail, which algorithms statistical learners converge on to solve linguistic tasks."

Nevertheless, because of the difficulty in interpreting not only neural networks but also interpretability results, we are perhaps more skeptical than Millière about how clean the answers will be as to which algorithms are the real ones. It's possible that we will have to get more comfortable with shades of gray when it comes to linguistic representation and algorithms. We see this work – figuring out the right methods for interpretability and then deploying them for linguistic and cognitive insights – as one of the most important areas for future work not just in linguistics-focused AI but in AI research more generally.

R5. Towards a more expansive LM research program

Many commentators criticize the current LM research program. In many cases, we fully agree with these criticisms. Where we disagree is treating present limitations as principled ceilings – what we call the As Good As It Gets Assumption.[3] This assumption proceeds by identifying a limitation of current LMs or of current linguistic research using LMs and then treats these as fundamental limitations. LMs are still in their infancy, and so is LM research. We see critiques as calls for *more* research into LMs to see whether these really are limitations. As such, what follows is what we see as a promising "yes, and" research program that acknowledges the limitations and then proposes ways to investigate them. Doing

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

exactly this kind of work is what we need in order to reveal whether these limitations will point to fundamental ways in which our central thesis is too optimistic.

R5.1. Learning: Data size, data diversity, and acquisition-relevance

We agree with Ambridge that "disruption is mutually assured" between LMs and research in language acquisition. Ambridge and a number of other commentaries draw our attention to data used to train models and how it compares (unfavorably in richness, although favorably in number of words) to the input a child gets. Lampinen highlights data diversity and the relative lack of diversity in internet text data compared to the kind of data children learn from. Wilcox and Newport call for a need for "human-sized data" and discuss the ways that LMs can help us understand how "graded, fuzzy constraints" interact during language acquisition, leading to a richly realized complex system. Howitt and Lidz and particularly McDermott-Hinman and Feiman discuss the ways that LM learning trajectories fundamentally differ from those of human learners.

Studying these aspects of how models learn is a promising research program. Indeed, we think efforts, inspired by child language learning, to train smaller models (Warstadt et al., 2023) under various manipulations (Misra & Mahowald, 2024; Patil et al., 2024; Yao et al., 2025) and study learning have been one of the most promising contributions of linguistics to LM research.[4] The commentators lay out promising avenues for further work in these domains. If it turns out that certain linguistic structures are fundamentally not amenable to being learned on human-scale data or that learning trajectories are fundamentally not recoverable by models, that will be an informative data point.

R5.2. Beyond standard written English: Multilinguality, dialects, interaction

Several commentators suggest that, by arguing for the role of LMs in linguistics, we privilege the role of standard written English, particularly syntax and semantics. Tripp writes, "LMs as understood in this frame effectively erase ways of languaging which are not readily captured in large repositories of text strings. Multimodal signals, admixtures of codes and the use of sign languages are all ignored, implicitly treated as complexity which is justifiably lost in message 'compression."

While much of this is a fair criticism of current LM research, the arguments we make in our target article also apply in principle to future models trained on non-standardized dialects, on audio, on sign language, on admixtures of codes, and on multimodal interactive systems. Model building can and should encompass these domains. And we see the body of work on written English not as the endpoint, but as a playbook for future work in all of these areas.

We do agree with Tripp that, in our review of the literature, we overfocused on syntax at the expense of other aspects of "languaging." We did so largely both because of the historical importance of syntax in linguistics and cognitive science and also because that's been the locus of most linguistic research on LMs (probably for the same reason: historical import). We enthusiastically endorse the pleas to push beyond the current boundaries. Below, we outline a few ways we think that can and will happen.

Non-English languages, like Tripp and Ármannsson et al., point to failures of LMs in other languages, particularly highlighting Icelandic. Modern LM architectures likely favor English (Blasi, Anastasopoulos, & Neubig, 2022), and more engineering effort has gone into English-language models. These are genuine shortcomings. We see expanding the scope of languages under study as one of the highest-priority items on the research agenda, both as an ethical imperative (since we want systems that work for everyone) and a scientific one (since we want to know how generalizable LM-based results are across languages and how typological differences interact with inductive biases, model performance, learning trajectories, and the like).

Having said that, we do not think that the current narrowness of LM-based results should *necessarily* be taken to make the extant body of LM work uninformative about language in general. For instance, consider the claim that English LMs deliver proof of concept that language *in general* can be learned by general-purpose learners with weak inductive biases. It would be scientifically surprising if English were learnable by large neural systems while typologically different languages were not learnable even in principle (even though other languages might benefit from different architectures or training regimes). A long-held tenet of linguistics is that all languages are learnable by children on roughly similar timescales. Thus, if English were uniquely amenable to statistical approaches, whereas other world languages were impervious to statistical learning, we would find that extremely surprising and thus extremely informative.

Audio We agree with de Heer Kloots et al. that linguists should also learn to love audio-based LMs. Doing so would extend the insights of this research program into the auditory domain, with import for phonetics and phonology. There is a lot of work left to be done towards this goal. At the same time, while text data lacks the multimodal richness of child input, we think it's sometimes understated how rich it is. While traditional corpora used in corpus linguistics came from news text and books, modern LMs are often trained on almost the whole internet. This includes all kinds of formal writing but also lots of informal writing: texts, web posts, and so on. In many cases, these data sources approximate informal real-time conversations, using emoji or other cues. So, while extending LM work to audio is an important future goal, we again do not think the extant literature is radically undermined by being text-focused.

Interaction and Communication Several commentators pointed out that there is something sterile about LMs: they lack the social, embodied, goal-laden usage that is the hallmark of human language. In that vein, we are sympathetic to much of Bunzeck et al.'s argument that LMs find patterns in data but are not fundamentally as "usage-based" as we say since they lack interaction. Culbertson et al. make a related point, drawing on the language evolution literature to argue that communication provides a crucial counteracting pressure that prevents language from collapsing into highly predictable but non-functional systems. The extent to which LMs are subject to these pressures is unclear. We find this perspective compelling and agree that the communicative function of language is essential to any complete explanatory account.

There is something off about merely learning from co-occurrence statistics, without having rich communicative goals. And it's true that classical pretrained models fall short in that way. But, as we move past the era of text-based pretrained models (as we largely have), we are optimistic that these limitations can be overcome.

As such, we see the lack of interaction and communication not as a strict limit, but as a demonstration that the right training regime matters. Regimes incorporating communicative objectives – whether through reinforcement learning from human feedback, multi-agent interaction, or grounded language tasks – are the path

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

forward. For instance, with the rise of AI agents, we expect to see a sharp increase in the amount of linguistics-focused work exploring AI behavior in much richer settings. While simple next-word predictors might not have goals or intentions of their own, an AI agent may well inhabit a setting that requires them to have goals (e.g., helping a user write successful code or complete a task). These agents may well have the kind of interactive, communicative setting that some commentators found lacking. Progress in AI is happening so fast that studying linguistic behavior only in non-communicative next-word prediction machines may soon look quaint.

In that vein, we share Sripada et al.'s optimism about the ability of LM-based linguistics to make contact with not just how an agent says something but also how it knows what to say when. They revive an old and somewhat dormant idea in linguistics: the Creative Aspect of Language Use (CALU), the ability of language users to select what they will say out of all the possible utterances they could say. We share Sripada et al.'s intuition that CALU has lain dormant, in part, because the problem was just too hard for the methods we had. And we agree modern AI models may change that. While there has been a lot of focus on text-based LMs, AI models in general can (and now often do) have visual components, information from various other modalities, and rich representations of the world. They are ready-made for grounding since all of these inputs can exist in a shared vector space, right along with the language. In that sense, "saying the right thing" can become a matter of saying the right thing conditioned on a particular non-linguistic state. We think this plausibly makes much richer modeling of choosing utterances "appropriate to the situation and linguistic context" far more tractable.

R6. Separate the science from the technology

Several commentators (Resnik, Tripp, Murphy et al., and Dingemanse and Cuskley) point out the irony of our target article's title, a reference to Kubrick's Atomic Age satire *Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb*. In our adaptation, "The Bomb" becomes "the Language Models." Some also noticed that another Kubrick film mentioned in our article, "2001: A Space Odyssey," in the introduction, is also an example where language technology (in the form of HAL) does not go particularly well for humans. We did not arrive at these Kubrick references with eyes wide shut. Indeed, we see the atomic metaphors as apt: the current moment in AI is a major technological moment that will significantly impact humanity in ways we can anticipate and ways we can't. While our take for what we can learn about human language on a scientific level is optimistic, we also see the potential for AI harm. We expect that a lot of human research and effort will be needed to make the AI Age "go well," much as there was massive effort required to make the Atomic Age "go well" and avoid catastrophe.

To that end, we urge language science to stay on the right side of Dingemanse and Cuskley's statement: "Robust research centers values over technology, moving from unexamined technopositivism to a concern with doing the best science possible." We advocate for LM research not out of unexamined technopositivism but exactly because we find it scientifically valuable. Dingemanse and Cuskley point to a number of shortcomings of existing work with AI, including the closed-source nature of commercial models and the lack of reproducibility. The academic community is well-positioned to tackle these challenges and continue to uphold scientific standards.

As such, we would refocus Dingemanse and Cuskley's argument on how the academic and nonprofit communities have risen to the very challenges they lay out. They worry about training data "ever since BERT" being a "carefully guarded and legally fraught secret," but now open-source models like Pythia (from EleutherAI, a non-profit) and OLMo (from the non-profit Allen Institute for AI) have fully transparent training pipelines. "Their frictionless design blinds us to underlying processes and transformations," and yet a thriving field of mechanistic interpretability focuses on understanding their underlying processes. "Their blackboxed nature endangers reproducible research," but efforts within the AI community have focused on replicating and making reproducible workflows. We don't see the project of studying LMs as trading one black box (brain) for another but as trading one black box for a system for which we know exactly what it was trained on, how its architecture is structured, and what kind of tools seem most promising for understanding its inner workings.

The project of understanding LMs is part of a broader scientific enterprise, one that has already been fruitful. The field of connectionism developed prototypes that eventually led to models that could produce apparently fluent language. The scientific community took these models apart, probed them, and asked whether they could *really* do what they seemed to. In some cases, they couldn't, and that led to improvements. In other cases, they could, and the question became how. This is instructive as a model for AI more generally. Just as LMs can inform our understanding of language and cognition, the scientific toolkit from linguistics and cognitive science can contribute to ongoing projects in AI interpretability and safety – projects where a robust, reproducible, reliable science is even more crucial.

We continue to believe that, as a field, linguistics has an important role to play in this enterprise. It is a gift to the field that a key to unlocking today's AI capacities has been the language model, an object steeped in linguistic and cognitive relevance. In many ways, we are optimistic that the field is meeting the occasion: a robust and healthy LM-infused science has emerged within many sub-areas of linguistics in the last several years. We hope to continue to learn about human language and human cognition from the most linguistically capable model humanity has ever built.

Acknowledgements. For helpful comments in preparing this draft, we thank Daniel Drucker. For comments on the draft, we thank Robbie Kubala, Kanishka Misra, and Chris Potts. KM acknowledges funding from NSF CAREER grant 2339729. Generative AI (Claude and ChatGPT) was used for brainstorming, summarizing, and searching for relevant literature.

Notes

1 We do not generally endorse the method of using one-off prompts to evaluate models, but we found it notable that Claude Opus 4.6 gave the intended answer from Murphy et al. and further added the caveat that it might differ if the argument structure is not like English "please": "If the subject/object roles reverse relative to English "please" (like how "like" and "please" work differently across languages — "me gusta" vs. "I like"), then which NPs fill which argument positions changes entirely." We judge this to be meeting and exceeding even the extremely high standard of metalinguistic competence set by Murphy et al. (2025).

2 Goodale and Mascarenhas acknowledge the need to connect to something beyond language by connecting semantics to symbolic theories of thought. Yet there is no more consensus on the discrete, symbolic, compositional nature of thought than there is for language. To make their point, they favorably cite Quilty-Dunn et al. on the Language of Thought (Quilty-Dunn et al., 2023). But Quilty-Dunn et al. discuss probabilistic Language of Thought in depth and approvingly, writing, "We're happy to let a thousand representational formats bloom."

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

3 An interesting historical example of this fallacy is in Pinker and Prince's (1988) famous argument against Rumelhart and McClelland's (1987) early connectionist work on the English past tense. Pinker and Prince (1988) focus a lot of their argument on criticizing the use of unordered phonetic features called "Wickelphones" as the basic unit in the model. This criticism is valid: it was, in hindsight, a strange methodological choice to use these kinds of features, a discarded idea unlikely to show up in any of the neural network work being done today. Based on the arguments made and the amount of ink devoted to Wickelphones, one gets the sense that it seemed to Pinker and Prince (1988) and many that it was a fundamental limitation of the whole connectionist program for language. But it wasn't; it was just a particular choice made in that particular implementation. When it came to neural implementations, Wickelphones were not as good as it got.

4 This line of work on small, custom-trained models is worth bearing in mind when considering Bolhuis et al.'s concern about the amount of data and energy needed to train models. They note that Google has ordered "six or seven small nuclear reactors for its AI data centers," but this conflates industrial-scale AI with the models actually used in linguistic research. Many models used in linguistics, including the ones mentioned above, involve training small models from scratch – running in hours on a modest GPU or even a laptop. A back-of-the-envelope calculation: training such a model at home costs around 36 cents in electricity, similar to baking a cake or doing laundry. The nuclear reactors just aren't a useful comparison for many of the models studied in the science of language.

References

Ahuja, K., Balachandran, V., Panwar, M., He, T., Smith, N. A., Goyal, N., & Tsvetkov, Y. (2025). Learning syntax without planting trees: Understanding hierarchical generalization in transformers. *Transactions of the Association for Computational Linguistics*, *13*, 121–141.

Allen-Zhu, Z., & Li, Y. (2025). Physics of language models: Part 1, Learning hierarchical language structures. *Transactions on Machine Learning Research*, *12*.

Baroni, M., Bernardi, R., & Zamparelli, R. (2014). Frege in space: A program for composition distributional semantics. *Linguistic Issues in Language Technology*, *9*, 241–346.

Berwick, R. C., Pietroski, P., Yankama, B., & Chomsky, N. (2011). Poverty of the stimulus revisited. *Cognitive Science*, *35*(7), 1207–1242.

Bhattamishra, S., Ahuja, K., & Goyal, N. (2020). On the practical ability of recurrent neural networks to recognize hierarchical languages. In D. Scott, N. Bel, & C. Zong (Eds.), *Proceedings of the 28th international conference on computational linguistics* (pp. 1481–1494). International Committee on Computational Linguistics.

Blasi, D., Anastasopoulos, A., & Neubig, G. (2022). Systematic inequalities in language technology performance across the world's languages. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 5486–5505). Association for Computational Linguistics.

Boguraev, S., Potts, C., & Mahowald, K. (2025). Causal interventions reveal shared structure across English filler–gap constructions. In C. Christodoulopoulos, T. Chakraborty, C. Rose, & V. Peng (Eds), *Proceedings of the 2025 conference on empirical methods in natural language processing* (pp. 25021–25042). Association for Computational Linguistics.

Brinkmann, J., Wendler, C., Bartelt, C., & Mueller, A. (2025). Large language models share representations of latent grammatical concepts across typologically diverse languages. In L. Chiruzzo, A. Ritter, & L. Wang (Eds.), *Proceedings of the 2025 conference of the nations of the Americas chapter of the association for computational linguistics: Human language technologies (volume 1: Long papers)* (pp. 6131–6150). Association for Computational Linguistics.

Cagnetta, F., Petrini, L., Tomasini, U. M., Favero, A., & Wyart, M. (2024). How deep neural networks learn compositional data: The random hierarchy model. *Physical Review X*, *14*(3), 031001.

Cagnetta, F., & Wyart, M. (2024). Towards a theory of how the structure of language is acquired by deep neural networks. *Advances in Neural Information Processing Systems*, *37*, 83119–83163.

Chaves, R. (2020). What don't RNN language models learn about filler-gap dependencies? In A. Ettinger, G. Jarosz, & J. Pater (Eds.), *Proceedings of the society for computation in linguistics 2020* (pp. 1–11). Association for Computational Linguistics.

Chomsky, N. (1957). *Syntactic structures*. Walter de Gruyter.

Chomsky, N. (1981). *Lectures on government and binding*. Foris Publications.

Clark, A. & Lappin, S. (2010a). *Linguistic nativism and the poverty of the stimulus*. John Wiley & Sons.

Clark, A. & Lappin, S. (2010b). Unsupervised learning and grammar induction. *The Handbook of Computational Linguistics and Natural Language Processing*, 197–220.

Coecke, B., Sadrzadeh, M., & Clark, S. (2010). Mathematical foundations for a compositional distributional model of meaning. Lambek Festschrift. *Linguistic Analysis*, *36*, 345–384.

Dennett, D. C. (1989). *The intentional stance*. MIT Press.

Eronen, M. I., & Bringmann, L. F. (2021). The theory crisis in psychology: How to move forward. *Perspectives on Psychological Science*, *16*(4), 779–788.

Futrell, R., Wilcox, E., Morita, T., Qian, P., Ballesteros, M., & Levy, R. (2019). Neural language models as psycholinguistic subjects: Representations of syntactic state. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 32–42). Association for Computational Linguistics.

Gauthier, J., Hu, J., Wilcox, E., Qian, P., & Levy, R. (2020). SyntaxGym: An online platform for targeted evaluation of language models. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 70–76.

Geiger, A., Harding, J., & Icard, T. (2025). How causal abstraction underpins computational explanation. arXiv preprint *arXiv:2508.11214*.

Geiger, A., Ibeling, D., Zur, A., Chaudhary, M., Chauhan, S., Huang, J., Arora, A., Wu, Z., Goodman, N., Potts, C., & Icard, T. (2025). Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, *26*(83), 1–64.

Gibson, E., & Wexler, K. (1994). Triggers. *Linguistic Inquiry*, *25*(3), 407–454.

Hendrycks, D., & Hiscott, L. (2025). The misguided quest for mechanistic AI interpretability. AI Frontiers. Guest Commentary.

Hewitt, J., Hahn, M., Ganguli, S., Liang, P., & Manning, C. D. (2020). RNNs can generate bounded hierarchical languages with optimal memory. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.) *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)* (pp. 1978–2010). Association for Computational Linguistics.

Hosseini, E., Casto, C., Zaslavsky, N., Conwell, C., Richardson, M., & Fedorenko, E. (2024). Universality of representation in biological and artificial neural networks. bioRxiv.

Hsu, A., Chater, N., & Vitányi, P. (2011). The probabilistic analysis of language acquisition: Theoretical, computational, and experimental analysis. *Cognition*, *120*(3), 380–390.

Hu, J., Mahowald, K., Lupyan, G., Ivanova, A., & Levy, R. (2024). Language models align with human judgments on key grammatical constructions. *Proceedings of the National Academy of Sciences*, *121*(36), e2400917121.

Hu, J., Wilcox, E. G., Song, S., Mahowald, K., & Levy, R. (2026). What can string probability tell us about grammaticality? *Transactions of the Association for Computational Linguistics*, *14*, 124–146.

Huh, M., Cheung, B., Wang, T., & Isola, P. (2024). The platonic representation hypothesis. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, & F. Berkenkamp (Eds.), *Proceedings of the 41st international conference on machine learning, volume 235 of Proceedings of Machine Learning Research* (pp. 20617–20642). PMLR.

Hunter, T., & Dyer, C. (2013). Distributions on minimalist grammar derivations. Proceedings of the 13th Meeting on the Mathematics of Language (MoL 13), pp. 1–11.

Kallini, J., Papadimitriou, I., Futrell, R., Mahowald, K., & Potts, C. (2024). Mission: Impossible language models. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 14691–14714). Association for Computational Linguistics.

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press

Kobzeva, A., Arehalli, S., Linzen, T., & Kush, D. (2025). Learning filler-gap dependencies with neural language models: Testing island sensitivity in Norwegian and English. *Journal of Memory and Language*, *144*, 104663.

Korsky, S. A., & Berwick, R. C. (2019). On the computational power of RNNs. arXiv preprint *arXiv:1906.06349*.

Lan, N., Chemla, E., & Katzir, R. (2024b). Large language models and the argument from the poverty of the stimulus. *Linguistic Inquiry*, 1–28.

Lan, N., Chemla, E., & Katzir, R. (2024a). Bridging the empirical-theoretical gap in neural network formal language learning using minimum description length. In L.-W. Ku, A. Martins, & V. Srikumar (Eds), *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, (pp. 13198–13210). Association for Computational Linguistics.

Lan, N., Geyer, M., Chemla, E., & Katzir, R. (2022). Minimum description length recurrent neural networks. *Transactions of the Association for Computational Linguistics*, *10*, 785–799.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, *101*(4), 676.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W.H. Freeman & Company.

McCoy, R. T., Frank, R., & Linzen, T. (2018). Revisiting the poverty of the stimulus: Hierarchical generalization without a hierarchical bias in recurrent neural networks. *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 40.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*(4), 806–834. https://doi.org/10.1037/0022-006X.46.4.806

Misra, K., & Mahowald, K. (2024). Language models learn rare phenomena from less rare phenomena: The case of the missing AANNs. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 913–929). Association for Computational Linguistics.

Mueller, A., Brinkmann, J., Li, M., Marks, S., Pal, K., Prakash, N., Rager, C., Sankaranarayanan, A., Sharma, A. S., Sun, J., Todd, E., Bau, D., & Belinkov, Y. (2026). The quest for the right mediator: Surveying mechanistic interpretability for NLP through the lens of causal mediation analysis. *Computational Linguistics*, *52*(1), 1–48.

Murphy, E., Leivada, E., Dentella, V., Montero, R., Günther, F., & Marcus, G. (2025). Fundamental principles of linguistic structure are not represented by ChatGPT. *Biolinguistics*, *19*, 1–55.

Musso, M., Moro, A., Glauche, V., Rijntjes, M., Reichenbach, J., Büchel, C., & Weiller, C. (2003). Broca's area and the language instinct. *Nature Neuroscience*, *6*(7), 774–781.

Núñez, R., Allen, M., Gao, R., Miller Rigoli, C., Relaford-Doyle, J., & Semenuks, A. (2019). What happened to cognitive science?. *Nature Human Behaviour*, *3*(8), 782–791.

Papadimitriou, I., Chi, E. A., Futrell, R., & Mahowald, K. (2021). Deep subjecthood: Higher-order grammatical features in multilingual BERT. In P. Merlo, J. Tiedemann, & R. Tsarfaty (Eds.), *Proceedings of the 16th conference of the European chapter of the association for computational linguistics: Main volume* (pp. 2522–2532). Association for Computational Linguistics.

Patil, A., Jumelet, J., Chiu, Y. Y., Lapastora, A., Shen, P., Wang, L., Willrich, C., & Steinert-Threlkeld, S. (2024). Filtered corpus training (FiCT) shows that language models can generalize from indirect evidence. *Transactions of the Association for Computational Linguistics*, *12*, 1597–1615.

Pearl, L., & Sprouse, J. (2013). Computational models of acquisition for islands. In J. Sprouse, & N. Hornstein (Eds.), *Experimental syntax and islands effects* (pp. 109–131). Cambridge University Press.

Perfors, A., Tenenbaum, J. B., & Regier, T. (2013). The learnability of abstract syntactic principles. *Cognition*, *118*(3), 306–338.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*(1–2), 73–193.

Quilty-Dunn, J., Porot, N., & Mandelbaum, E. (2023). The best game in town: The reemergence of the language-of-thought hypothesis across the cognitive sciences. *Behavioral and Brain Sciences*, *46*, e261.

Rumelhart, D. E., & McClelland, J. L. (1987). Learning the past tenses of English verbs: Implicit rules or parallel distributed processing? In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 195–248). Lawrence Erlbaum Associates.

Ryu, S. H., & Lewis, R. L. (2025). Memory for prediction: A transformer-based theory of sentence processing. *Journal of Memory and Language*, *145*, 104670.

Simonite, T. (2018). Google's AI guru wants computers to think more like brains. WIRED. Accessed 2026-02-11.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In D. Yarowsky, T. Baldwin, A. Korhonen, K. Livescu, & S. Bethard (Eds.), *Proceedings of the 2013 conference on empirical methods in natural language processing* (pp. 1631–1642). Association for Computational Linguistics.

Spivey, M., & Tanenhaus, M. (1998). Syntactic ambiguity resolution in discourse: Modeling the effects of referential context and lexical frequency. *Journal of Experimental Psychology, Learning, Memory and Cognition*, *24*, 1521–1543.

Sutter, D., Minder, J., Hofmann, T., & Pimentel, T. (2025). The non-linear representation dilemma: Is causal abstraction enough for mechanistic interpretability? In *39th Conference on Neural Information Processing Systems (NeurIPS 2025)*.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634.

Trueswell, J. C., Tanenhaus, M. K., & Garnsey, S. M. (1994). Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Memory and Language*, *33*, 285–318.

Warstadt, A., Mueller, A., Choshen, L., Wilcox, E., Zhuang, C., Ciro, J., Mosquera, R., Paranjabe, B., Williams, A., Linzen, T., & Cotterell, R. (2023). Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*.

Weiss, G., Goldberg, Y., & Yahav, E. (2018). On the practical computational power of finite precision RNNs for language recognition. In I. Gurevych, & Y. Miyao (Eds.), *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 740–745). Association for Computational Linguistics.

Wilcox, E., Levy, R., Morita, T., & Futrell, R. (2018). What do RNN language models learn about filler–gap dependencies? In T. Linzen, G. Chrupala, & A. Alishahi (Eds.), *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP* (pp. 211–221). Association for Computational Linguistics.

Wilcox, E. G., Futrell, R., & Levy, R. (2024). Using computational models to test syntactic learnability. *Linguistic Inquiry*, *55*(4), 805–848.

Wilson, A. G. (2025). Position: Deep learning is not so mysterious or different. *Proceedings of the 42nd International Conference on Machine Learning*, Vancouver, Canada.

Yao, Q., Misra, K., Weissweiler, L., & Mahowald, K. (2025). Both direct and indirect evidence contribute to dative alternation preferences in language models. *COLM 2025*.

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, *64*(3), 107–115.

https://doi.org/10.1017/S0140525X2510112X Published online by Cambridge University Press